ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

PREDICTING HIV INFECTION RISK FACTOR USING VOLANTORY COUNSELING AND TESTING DATA: A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)

GIRMA AWEKE

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

PREDICTING HIV INFECTION RISK FACTOR USING VOLANTORY COUNSELING AND TESTING DATA:

A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information science

By

GIRMA AWEKE

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

PREDICTING HIV INFECTION RISK FACTOR USING VOLANTORY COUNSELING AND TESTING DATA: A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)

By

GIRMA AWEKE

Name and signature of members of the examining Board

Name	**Title**	**Signature**	**Date**
W/ Lemlem Hagos	Chairperson	_____	_____
Dr. Dereje Teferi	Advisor	_____	_____
Dr Rahel Bekele	Examiner	_____	_____

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Girma Aweke

JUNE 2012

The thesis has been submitted for examination with my approval as university Advisor

---

Dr. Dereje Teferi

JUNE 2012

DEDICATION

I would like to dedicate this thesis work to my uncle Abera Dinkayehu who has been struggling and fighting for my education since I was a little boy. Gashe you are always in my heart and your dream is now realized!!

ACKNOWLEDGMENT

First and foremost extraordinary thanks go for my Almighty God and His Mother Saint-Marry.

I would like to express my gratitude and heartfelt thanks to my advisor, Dr. Dereje Teferi for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study.

I am very grateful to the management and staff of AAII, especially Ato Mollalign, for constant support in accessing the data and Ato Zerihun for providing appropriate professional comment, explanation and other supportive document for my research work. My special thanks also goes to Addis Ababa University Library Management for granting me study leave with the necessary benefits, without which I could not have been able to join my M.Sc. study.

I am immensely indebted to my beloved family especially My wife w/o Asnakech Simegn ,my lovely daughter Elishaday Girma, My father Aweke D.,My Mother Almaz N., My Brother Demoz, My sisters Bethlehem, Gete and not mentioned here parents for giving me unconditional care, love, time, patient and support throughout my life.

My special thanks also goes to my friend and classmate Biniam Asnake.Your support and advice has been worth considering starting from the beginning up to the completion of the program.

I also grateful to my friends, staff members of school of Information science(especially Dr. Million Meshesha and other), and my classmate (Tigabu, sewale,Negasi,G/medin,Biazen,Danil M. and other) for encouragement, support and the good friendship they shared.

Last, not least special thanks also goes to Ato Adane Letta and Tesfaye H/mariam(for editing and providing essential comment) , W/t Meseret Ayano for her unlimited support of all relevant information resources whenever her cooperation was needed.

My special thanks also goes to Ato Solomon M , Nebyou A., Abraham G.,Ato Mesfin G., Ato Belay, and Ato Behailu J. who offered good hospitality whenever their cooperation were needed.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAII	African AIDS Initiatives International
AAU	Addis Ababa University
AIDS	Acquired Immunodeficiency Syndrome
ARFF	Attribute-Relation File Format
CDC	Center of Disease Control and Prevention
CHAID	Chi-squared Automatic Interaction Detection
CART	Classification and Regression Trees
CRISP-DM	CRoss-Industry Standard Process -Data Mining
CSV	Comma Separated Values
DM	**Data Mining**
HAPCO	HIV/AIDS Prevention and Control Office
HIV	Human Immunodeficiency Virus
KDD	Knowledge Discovery in Database
MoH	Ministry of Health
UNAIDS	United Nations AIDS
UNFPA	United Nations Population Fund
VCT	Voluntary Counselling and Testing
WHO	World Health Organization
WEKA	Waikato Environment for Knowledge Analysis

Table of Contents

ABSTRACT

Despite a great deal of efforts, the world still has neither a cure nor a vaccine for HIV/AIDS infection. Millions of people have been suffering from this incurable disease. Fortunately, researchers have become successful in prolonging and improving the quality of life of those infected with HIV. Nonetheless, it has become increasingly clear that preventing the transmission and the acquisition of HIV through educating people to bring about behavioral changes should be the focus.

The widely and freely available voluntary counseling and testing center (VCT) in Addis Ababa which provides an enormous role in counseling, promoting and checking clients HIV status through the clinical laboratory test. In line with this, the center has been collecting client's information or records for further investigation with confidentiality. The record consists of many attribute that may have a direct or indirect impact with HIV infection. Moreover, identifying HIV infection risk factors or determinate variables provides benefits at different level of the society (such as individual, community and organizational level).

The benefit not yet known by the client rather the organization keeps their records after they got tested. To this end, great efforts have been made to develop models to identify HIV infection risk factor using data mining technology.

This research is initiated to identify the determinant risk factors of HIV infection by developing predictive models to support voluntary counselling and testing service of African AIDS Initiatives international (AAII) provided at Addis Ababa University and its surrounding.

The six steps hybrid methodology has been followed for predictive HIV infection risk factors modeling among selected attributes. Three classifications techniques such as Decision tree J48, PART and SMO algorithms were experimented for building and evaluating the models.

Before experimentation data pre-processing task has been performed to remove outliers, fill in missing values, and select best attributes, discretization and transformation of data. The pre-processing phases took considerable time of this work. A total of 15,396 VCT client records have been used for training the models, while a separate 3,000 records were used for testing their performance. The model developed using the PART algorithm has shown the best classification accuracy of 96.7%. The model has been evaluated on the testing dataset and scores a prediction accuracy of 95.8%. The results of this study have shown that the data mining techniques were valuable for predicting HIV infection risk factors. Hence, future research directions are forwarded to come up applicable solutions in the area of the study.

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Despite a great deal of efforts, the world still has neither a cure nor a vaccine to HIV/AIDS (Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome) infection. Millions of people are suffering from this incurable disease. Fortunately, researchers have become successful in prolonging and improving the quality of life of those infected with HIV. Nonetheless, it has become increasingly clear that preventing the transmission and the acquisition of HIV through educating people to bring about behavioral changes should be the focus (UNAIDS, 2010).

Although the United Nations Aids Department (UNAID) has reported that the HIV prevalence is leveling off and there is a fall in the numbers of new infections globally (UNAIDS, 2010), it remains one of the leading obstacles to health and growth of developing countries. There are still a huge number of people infected and affected by HIV. For instance, in 2009, there were an estimated 2.6 million people who became newly infected with HIV. The incidence of HIV infection declined by 19% between 1999 and 2009 globally; the decline exceeded 25% in 33 countries, including 22 countries in sub-Saharan Africa. However, it was also reported that HIV/AIDS has become the leading cause of death in the world (UNAIDS, 2010).

Ethiopia is one of the Sub-Saharan African countries most severely affected by the HIV/AIDS pandemic. Currently, the national adult prevalence rate is estimated at 2.3 percent and an estimated number of 1.2 million people are living with HIV/AIDS , an estimated 67,000 people lost their lives due to AIDS at the end of 2007 (UNAIDS/WHO, 2010). The Federal HIV/AIDS prevention and control office also report that young women were particularly vulnerable to HIV infection as compared to young men. Among women aged 15-19 and 20-24, for example, 0.7 and 1.7 percent, respectively, were HIV-infected, while the figure for men in the same age categories showed that only 0.1 and 0.4, respectively, were infected (FDRE HIV/AIDS PC, 2010).

Despite the overwhelming challenges, the Government of Ethiopia is making efforts towards containing the epidemic and mitigating the impact of HIV/AIDS through an intensified national response in a comprehensive and accelerated manner.

To this end, huge effort had been made to build the implementation capacity, especially in the health sector, in areas of human resource development, site expansion and construction of health facilities. There are various non- profit organizations that have the mission of preventing HIV infection, improve care and support and build capacity to address the global HIV/AIDS pandemic. Among them, African AIDS Initiative International (AAII) is the leading international organization founded on April 2001 officially by opening a regional office in Addis Ababa, Ethiopia (LLGP, 2007).

The office plays an enormous role in preventing the spread of HIV/ AIDS in the university community, in this regard, the statistics about Voluntary Counseled and Tested (VCT) service of AAU indicates that the number of treatment seekers and counseling test services has highly increased and currently estimated 18,800 clients have taken the service. Based on a recent research report of Nigatu and Seman (2011) out of the total respondents, 47.2% had been tested for HIV/AIDS and more than 80% had ingress to take VCT service for HIV/AIDS.

Additionally, the attitudes and practices of the Addis Ababa University community towards HIV prevalence has been exposed by the risk factors such as sex, previous place of residence, religious participation, pornographic film show, alcohol intake, chat chewing, and Cigarette smoking. A comprehensive approach to prevent HIV includes, creating an opportunity for people to participate in VCT service (Nigatu and Seman , 2011). Besides counseling and testing service, it is best to investigate the degree of association between the risk factor mentioned above which is equally important in eradicating HIV infection.

1.1.1. Data Mining and Health care

Widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer-assisted analysis (Larvac, 1998). Extensive amounts of data gathered in health care databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. It has been estimated that an acute care hospital may generate five terabytes of data a year (Fayyad et al., 1996). The ability to use these data to extract useful information for quality health care is crucial.

Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified; efficiency and accuracy of decisions decrease when humans are put into stress and immense work (Eapen , 2004). Prather et al. (2001) wrote that there are only a few tools to evaluate and analyze clinical data after it has been captured and stored. The researcher further stated that evaluation of stored clinical data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management.

The traditional method of turning data into knowledge relies on manual analysis and interpretation (Fayyad et al., 1996). Prather et al. (2001) argued that to evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data, to discover new patterns and relationships hidden in the data. It is due to these challenges of searching for knowledge of relational databases and our inability to interpret and digest these data as readily as they are accumulated, which has created a need for a new generation of tools and techniques for automated and intelligent database analysis. Consequently, the discipline of knowledge discovery in databases (KDD), which deals with the study of such tools and techniques, has evolved into an important active area of research (Raghavan et al., 2002).

Data mining is one among the most important steps in the knowledge discovery process. It can be considered the heart of the KDD process. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the data (Eapen, 2004). It is a new generation of computerized methods for extracting previously unknown, valid, and actionable information from large volume of database and then using this information to make critical decision (Cabena et al., 1998). Hence, it is better to explore the relevance of this technology to any organization that performs health-related activity.

There are various techniques and tools of data mining that help researcher in exploring the benefit of data mining in various sectors. Data mining techniques can be broadly classified based on what they can do, namely description and visualization; association and clustering; and classification and estimation, which is predictive modeling. Predictive modeling can be used to identify patterns, which can then be used to predict the odds of a particular outcome based upon the observed data.

Rule induction is the process of extracting useful if/then rules from data based on statistical significance. A decision tree is a tree-shaped structure that visually describes a set of rules that caused a decision to be made. For example, it can help determine the factors that affect kidney transplant survival rates and the nearest-neighbor method (DMSHO, 2011). Each of these techniques analyzes data in different ways.

In addition to data mining technique mentioned above, most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools (Silltow, 2006). According to Kon and Geraled (2005) there is vast potential for data mining applications in healthcare. Generally, these can be grouped as the evaluation of treatment effectiveness; management of health care; customer relationship management; detection of fraud and abuse; and the identification of risk factors associated with the disease.

1.2. Statement of the Problem

The costs of fatalities and severity due to HIV/AIDS have a tremendous impact on societal well-being and socioeconomic development. HIV/ADIS is among the leading causes of death worldwide, causing an estimated 1.8 million deaths per year (UNAIDS, 2010). According to the report of the second round HIV/AIDS Behavioral Surveillance Survey in Ethiopia, it was found out that around 9.9 percent of the in-school youth (14.6 % of males and 5.3 % of females) had sexual experience (BSS, 2005). Nigatu and Seman (2011) in their research reported that HIV/AIDS is affecting young members of the societies especially adolescents between the age of 15 to 24 and proved that these age groups were vulnerable and at risk of the disease.

A number of studies have also shown that AIDS has progressively been on the increase and constitutes a big problem among college and university students, although the extent of the problem is relatively unknown (Abdinasir and Sentayehu, 2002; Elias, 2009; Tefera, Challi and Yoseph, 2004.; Teka, 1993; Getinet, 2009; AAII, 2006). Evidence showed that most sexual risk behaviors among college and university students might have been acquired through a period of campus life (Teka, 1993). This may be due to the life of independence, away from parental control, that often characterizes such a setting.

University students are often viewed as being at high risk for HIV infection due to their propensity to engage in exploratory behavior and their needs for peer social approval and false sense of non-vulnerability (Beyene, Solomon and Yared, 1997).

A number of factors contribute to the cause of HIV infection such as economic, social, behavioral, environmental, educational and health related factors (Nigatu and Seman , 2011). Among these there are variable or attribute that have trivial and non-trivial role in causing a client being infected with HIV. Hence, categorizing and/or predicting HIV infection risk factor of attribute or variable that have causal relation or pattern from voluntary counseling and testing data is very crucial for organizations who are investing large amounts of money on preventing and controlling HIV/AIDS.

In attaining its objective, AAII collects daily, monthly, quarterly and yearly health-related statistics of the university community (LLGP, 2007). It analyzes the available data with traditional statistical tools or technique alone is not enough for health professional, planner and policy maker to identify major determinant risk factors for HIV infection and transmission. As Plate et al. (1997) indicated that traditional method of data analysis has limited capacity to discover new and unanticipated patterns and relationship that are hidden in conventional databases. Identifying patterns of attribute or risk factors of HIV infection is difficult because the dataset contains or involves too many attribute or parameter. So studying the patterns of the attribute that have relationship among each other rather than listing are more crucial.

To the knowledge of the researcher, no attempt has been made using rule induction and support vector machine classification technique to identify the determinant factor of HIV infection. Moreover, this work has been the first attempt to evaluate the applicability of DM for predicting HIV infection risk factor in higher education institutions. Last not least, three unique variables (Category, campus and Year of the student) have been used as an additional attributes for analysis of the experimentation that was not used in the previous studies.

Therefore, the purpose of this study is to investigate the underlying determinant risk factors of HIV infection from the available VCT data using data mining techniques. To this end, it has been attempted to obtain answer for the following research questions:

- What are the main determinate risk factors (attributes) of Voluntary counseled and tested client records that cause HIV infection in the Addis Ababa University community and their surroundings?
- What are the most interesting patterns or rules generated/predicted to determinate risk factor for HIV infection that can be used as a cause of HIV /AIDS.
- To what degree (accuracy level) the determinate risk factor for HIV infection be determined by applying data mining.

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this study is to apply data mining technology to identify and predict the major risk factor of HIV infection using VCT records by developing a model that can support counseling and testing service provider, policy makers, and planners.

1.3.2. Specific Objectives

For the realization of the general objective stated above, the following specific objectives have been formulated.

- To review the literature on DM technology and their application in the Health care Industry particularly in HIVAIDS for the purpose of getting information that will help in this research
- To prepare the data for analysis and model building, by cleaning and transforming the data into a format suitable for the selected DM algorithms
- To identify the features of risk factor for HIV infection, and select the appropriate classification algorithms to be used based on the type of data and objectives of the study
- To train and develop a classification model that support in predicting HIV infection risk factor
- To test and compare the resulting performances of the classification models and recommend the overall best results of the classification models.
- To report findings of the result and forward recommendation for further research.

1.4. Research Methodology

According to Smolander et al. (1990) a method can be considered as a predefined and organized collection of techniques and a set of rules which state by whom, in what order, and in what way the techniques are used to achieve or maintain some objectives. The DM process model describes procedures that are performed in each of its steps (Jinhong et al. 2009). It is primarily used to plan, work through, and reduce the cost of any given project.

1.4.1. Research Design

This research is designed to identify the determinate risk factor of HIV infection. To explore the application of data mining on this particular research, hybrid (Ciso.et al) data mining methodology was employed. This model was developed, by adopting the CRISP-DM model to the needs of academic research community (Fayyad, 1996). Unlike the CRISP-DM process model, which is fully industrial, the hybrid process model is both academic and industrial. Additionally, it extends it functionality in to research oriented description of the steps; incorporating DM step instead of the modeling step, and increasing number iteration among steps to make the research process more complete or to come up with explicit feedback mechanisms.

Ciso.et al involves sixth iterative process or steps including: understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge steps (Pete, 2000).

As depicted in figure 1.1 to determine HIV infection risk factor form VCT dataset data mining tools is not the only tools that we require instead review of literature(conceptual and theoretical) be done. Moreover, we need to understand the system that contain the initial data and other intermediate data processing tools and technique Such as MS Excel and Epi Info software to transfer and Pre-process the data for actual application of data mining task using Hybrid data mining model. The researcher tries to discuss each step in details below.

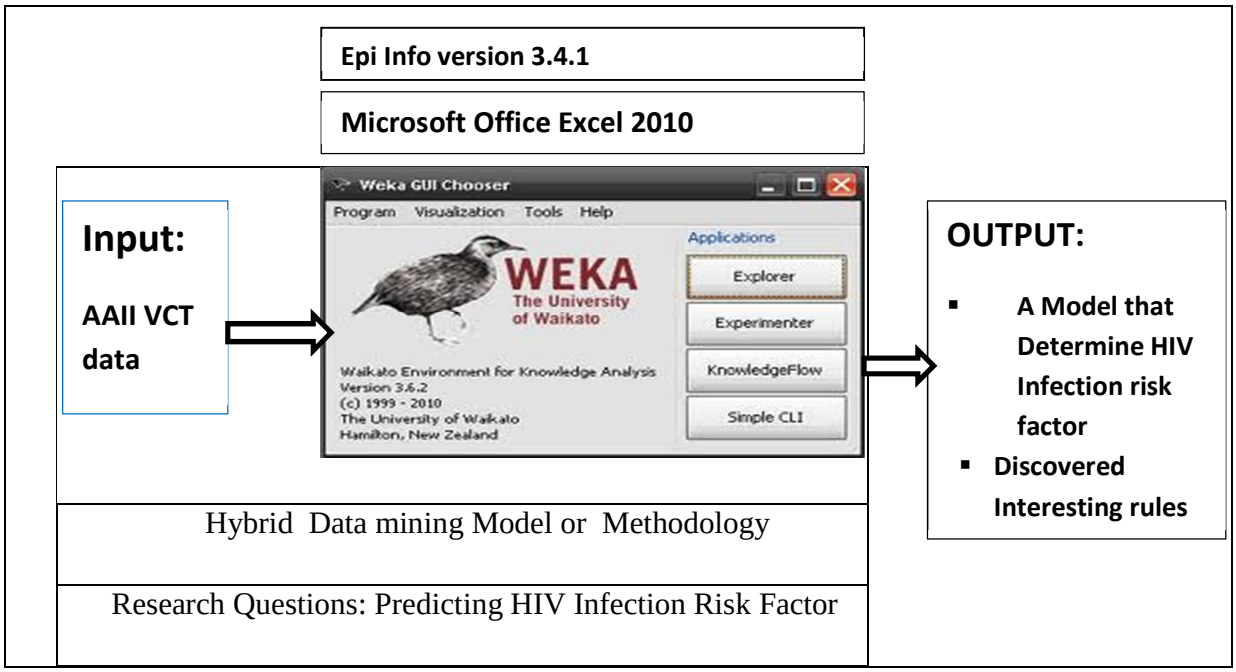


Figure 1. 1 Diagrammatic Skeleton of overall Research Methodology and Design

1.4.1.1. Understanding of the problem domain

The primary task of the researcher in knowledge discovery process is a close look of the problem domain. To identify, define, understand and formulate the problem domain (VCT service problems), various data collection methods were used including: interview, observation, document analysis. In-depth interview has been conducted with domain expert to determine attribute feature selection and understand some complex business process. As a result of insight gained knowledge of the domain data mining problem is defined. Further, databases such as VCT data, major service AAII provides and HIV/AIDS related policy and related issue are consulted to gather the pertinent data for the present research.

The main purpose of the AAII VCT centers is to save the life of the university community and surrounding particularly the industrious people by mitigating the prevalence of HIV/AIDS through establishing a permanent VCT center in each campus and surroundings.

1.4.1.2. Understanding the VCT data

Once the problem to be addressed is understood, the next step is analyzing and understanding the data itself. Because, the end result of knowledge discovery process heavily depends on the quality and quantity of available data (Cios et al., 2007). In this phase, the original data collected for this research should be described briefly. Its description includes listing out attributes with their respective values, missing and outlier and evaluation of their importance to the research goal.

Additionally, careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the DM tasks to be performed.

1.4.1.3. Preparation of the data

This is the most crucial phases in which the success of the entire knowledge discovery process depends. It usually consumes much of the entire research effort. In this phase, the final dataset is also constructed from the initial raw data. Data preparation tasks were performed in iterative ways. The major tasks include: description of data sources, carrying out statistical summary measure, finding out distinct value, filling missing values, outlier and noisy data and data transformation/reduction activities were also undertaken in this phase. Additionally, feature / attribute construction was also made too. The researcher decides, together with domain experts, the data with respective attribute that is used as input for applying the DM techniques (see Appendix 2).

Data cleaning (or data cleansing) routines was applied to fill in missing values (with the most frequented or modal value), smooth out noise (by removing the record), and detect outliers (by removing or substituting with moral values) in the data. To support the preprocessing tasks, the researcher used WEKA 3.7.5 to normalize and to fill missing value with modal value. It also helps to train the selected algorithms with training sample and testing case.

1.4.1.4. Data mining

The prerequisite for the application of data mining research goal is preprocessing task. Besides, better understanding and preparation of data result in a good selection of data mining methods

otherwise it leads the miner to incorrect decision (Fayyad, 1996). So this step is all about selecting appropriate data mining methods based on the research objectives.

The use of DM to gain knowledge discovery regularities in the data and not just predictions has been common (Witten and Frank, 2000). We can understand from the objectives of the research, gaining knowledge and discovering patterns that cause HIV infection within VCT data is certainly the purpose of the study. Consequently, a classification technique particularly decision tree, rule induction and support vector machine techniques were selected and used for prediction of the HIV infection risk factor. The researcher selected only classification technique because the research dataset has clear and simplified labeled class. For training and implementation purpose J48, PART and SMO algorithms were used respectively.

1.4.1.5. Analysis and Evaluation of the discovered knowledge

In data mining evaluation has two primary functions. Primarily, it helps to predict how well the final model will work in the future. Secondly, evaluation is an integral part of many learning methods and helps to explore the model that best represents the training data. According to Ciso et al. (2007) the novelty and interestingness of the models are evaluated with the agreement of the researcher and domain expert.

The different classification models developed in this research were evaluated using a test dataset based on their classification accuracy. In the other word evaluation of the performance of the classifier is also made in terms of different confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), Relative Operating Characteristics (ROC)), the number of correctly classified instances, number of leaves and the size of the trees, execution time.

Furthermore, models can be compared with respect to their speed, robustness, scalability, and interpretability which may have an influence on the model (Kantardzic, 2003). The confusion matrix of the classifier model was analyzed in terms of the following variables in Table 1.1

Table 1. 1 Sample summary of Confusion Matrix of the model

		Predicted the status of HIV infection		Total
		Negative	Positive	
Actual HIV Status	Negative	TN	FP	TN + FP
	Positive	FN	TP	TP + FN
Total		TN + FN	FP+ TP	TN + FP + FN + TP

Key: TN = True Negative FP= False Positive FN= False Positive TP= True Positive

Contextually what each variable stands for is stated as follows, TN: The number of HIV negative clients that are classified as negative. FN: The number of HIV positive clients that are classified as negative. TP: The number of HIV positive clients that are classified as positive. FP: The number of HIV negative clients that are classified as positive. FP + TN: The total number of actual HIV negative clients. TP + FN: The total number of actually HIV positive clients. TN + FN: The total number of predicted HIV negative clients. FP+ TP: The total number of predicted HIV positive clients. TN + FP + FN + TP: The total number of all clients.

The performance of the experiments is measured in the following manner as depicted below:

Sensitivity (TPR): indicated by the proportion of HIV positive clients that are correctly classified as positive i.e. TPR= TP/ TP + FN

Specificity (TNR): The proportion of HIV negative clients that are correctly classified as negative. TNR= TN/ TN+FN

False Negative Rate (FNR): The proportion of HIV positive clients that are erroneously classified as negative i.e. FNR= 1 – TPR or FNR = FN/ TP + FN hence, the sum of TPR plus FPR equals 1.

Correctly Classified Instances (Accuracy): To compute the proportion of clients those are correctly classified using this formula i.e. **Accuracy** = (TP + TN) / (TN + FP + FN + TP)
Similarly, Incorrectly Classified Instances (Error Rate): The proportion of clients that are incorrectly classified. Error Rate = (FP + FN) / (TN + FP + FN + TP).

Furthermore we can also compute the effectiveness and efficiency of the model in terms of recall and precision. As a result, recall can be computed for both positive and negative classes. So the formula of the recall for negative class be TN/TN + FP and for positive class TP/ TP+FN. Similarly to compute the perception of the model the following formula be used for negative and positive class respectively, TN/ TN + FN and TP/ FP+ TP.

Besides, this research plans to use the following methodology to develop data mining model that determines the risk factor for HIV infection.

1.4.2. Literature review

An in depth literature review is conducted by refereeing of books, journal articles, conference paper and the internet get more insight to the concept of data mining and its application, especially in health related areas.

1.4.3. Data identification, collection and preparation

The primary source of data to conduct this research was voluntary counseled and tested database of AAII which contains more than 18,800 records. To this end, the data are collected and arranged into new database to make it suitable for the experiment and for the selected data mining tools. Therefore, a new database was prepared by analyzing the collected data using preprocessing tasks like data cleaning and reduction.

1.4.4. Model building

As indicated in the research objective the researcher used data mining technique to develop a model that identify determinant and predict risk factors for HIV infection. So, in the experimental part predictive models has been developed by using the selected data mining technique. The

research was conducted with three selected techniques of data mining. Finally, a comparison of them was made to get a reasonable accuracy.

1.4.5. Tools and Software

The following tools and application software have been used to accomplish the research process: **WEKA 3.7.5 data mining tools**: Based on the prior knowledge of the researcher on the applicability of the tools for the purpose and freely availability of the software made the researcher to use the tools to build analysis and evaluate the model being developed for research goal.

Ms-Excel: it was used for data preparation, pre-processing and analysis task because it has the capability of filtering attribute with different values. Besides, it is a very important application software to make ready the data and easily convert into of the file.

Ms-word and Ms Power point: It is a known fact that the final report document is prepared with word and for presentation purpose preparing slide is inevitable. So the use of power point is mandatory.

Epi info: this is standard software that used to handle health related data. As a result, AAII handles its VCT data with it. So to pre-process the original data found from it the researcher employed this software specifically to calculate missing value and detect outlier.

1.5. Scope and Limitation of the Study

The main aim of this research is to find out the applicability of DM for determining and predicting HIV infection risk factor in the VCT provider center in Addis Ababa. This research focuses only on client records of Addis Ababa University (AAU) community and its surrounding (Arada sub-city) for the reason that university community is more exposed to those places to HIV/AIDS. Specifically, 15,396 records of data for knowledge discovery are obtained from the African AIDS Initiatives International (AAII).

The dataset of AAII also contains other higher education institution's client records. However, it is not included in this research. Among the 63 attributes, 20 of them are used in experimentation, which are selected by the domain experts.

In this study, we evaluated three classification techniques to develop a prediction model to identify major determinant risk factors for HIV infection.

1.6. Significance of the Study

Though the primary goal and initiatives of this research has been for academic exercise, the findings of this research can ultimately use in various areas.

The organization (both AAII and AAU) may use the predictive model to determine risk factor of HIV infection based on the available VCT data and understand the potential of each variable towards causing HIV infection in higher educational institutes particularly in Addis Ababa University. This helps to design and develop strategies for the efforts towards struggling against HIV/AIDS care, control and prevention program. To this end, the organization protects and save the lives of thousands of peoples.

The government, Ministry of Health, Ministry of education, the university management and policy maker can also use the output of the research as a decision support for formulating policy related to HIV care, control and prevention Program in higher education institutes.

The research also gave opportunity for utilization of health care data particularly VCT test data that is stored for no purpose , and the model developed also enables less qualified data manger or health professional staff assigned to perform the task of determining and managing HIV infection more accurately and confidently and less subjectively.

Last not least, the output of the research also contribute to understanding of the theoretical views and practical problems in the design and implementation of VCT programs and invite interested researchers to explore more in related and similar areas.

1.7. Organization of Thesis

The research report in this study has six chapters. The first chapter deals with the basic overview including background, statement of the problem, objective, scope of the study, methodology and thesis organization.

In the second chapter critical literature review which includes review of national and global research and trend of HIV/AIDS in higher education. Definition , benefit of VCT services has been discussed.

In the third chapters review of data mining literatures was presented. Definition, evolution, types, application of data mining in different domain areas was discussed. Review of related work particularly application of data mining in healthcare sector especially with the issue of HIV/AIDS was discussed here.

The fourth Chapter deals with data preprocessing tasks. In this chapter how the major data preprocessing tasks were applied to the current data were shown. Data cleaning, reduction and preparation of dataset to be used as input for predictive model.

The fifth chapter deals with experimentations and result interpretations. In this chapter building of model with training dataset and validating the result with testing datasets ,and interpretation of the result of the experimentation were the major concern. Finally, a comparison of the algorithms used for reasonable accuracy was made.

In the sixth chapter conclusions and recommendations were presented

CHAPTER TWO

HIV/AIDS AND CURRENT RESEARCH OUTPUT

Under this chapter attempt has been made to present a review of the literature and trends of HIV/AIDS globally as well as locally, overview, process and benefit review of voluntary counseling and testing (VCT) services in Ethiopia is also made. Lastly, application of data mining in healthcare particularly in HIV/AIDS is thoroughly reviewed. These help a great deal in understanding the domain area where data mining technology is going to be used.

2.1. Overview of HIV/AIDS

HIV, the virus that causes AIDS (Acquired Immunodeficiency Syndrome) (UNAIDSb., 2010) has become one of the world's most serious healthiest and development challenges. The first cases were reported in 1981 (AIDS, n.d.). Since then, the disease has spread rapidly across the globe; in 2010 sub-Saharan Africa was the most HIV infected region in the world with nearly 22.9 million (67%) people living with HIV in the world which is nearly two-thirds of the global burden [UNAIDSa ,2011).In Africa as a whole, the death rate in 2010 is reported to be 1.2 million [UNAIDS ,2011).

These figures are more staggering when you look at the results alongside findings that only 12% of men and 10% of women in the general population had been tested for HIV and received their results (WHOa, 2010). In addition, many of the men and women who seek HIV testing and counseling, are already in the advanced stages of the disease [Hogg, et al. 2006]. The WHO and UNAIDS proclaim that "early diagnosis presents an opportunity to provide people with HIV the information and tools to prevent HIV transmission to others" (WHOb,2004).

A voluntary counseling and testing (VCT) program has been formulated by the WHO and UNAIDS to encourage people to be pre-HIV test counseled, tested and post-HIV test counseled in an endeavor to prevent infection and transmission of HIV.

2.2. Global Issue of HIV/AIDS Epidemic

Global HIV/AIDS data show that the HIV/AIDS epidemic is a worldwide public health problem of great magnitude. About two decades since the beginning of the HIV/AIDS epidemic, over 34 million people worldwide have currently living with HIV and nearly 30 million people have died of AIDS-related causes since the beginning of the epidemic (UNAIDSa, 2011).

Although cases have been reported in all regions of the world, almost all those living with HIV (97%) reside in low- and middle-income countries, particularly in sub-Saharan Africa(UNAIDSa ,2011). Most people living with HIV or at risk for HIV do not have access to prevention, care, and treatment, and there is still no cure (WHOa,2010)

HIV primarily affects those in their most productive years; about half of new infections are among those under age 25 (UNAIDSa,2011). HIV not only affects the health of individuals, it impacts households, communities, and the development and economic growth of nations. Many of the countries hardest hit by HIV also suffer from other infectious diseases, food insecurity, and other serious problems. An evidence of this that people living with HIV at the end of 2010, up from 28.6 in 2001 to 34 million (UNAIDSc, 2011).

Despite these challenges, new global efforts have been mounted to address the epidemic, particularly in the last decade, and there are signs that the epidemic may be changing course. According to the Global HIV/AIDS Epidemic report (UNAIDSa, 2011) depicted in Table 2.1, even though, people living with HIV increase globally, deaths have declined due to the scale up of antiretroviral treatment (ART).

Besides that, Table 2.1 indicates that the global prevalence rate (the percent of people whose ages ranging from (15–40) are infected) was 0.8% in 2010 (UNAIDSa ,2011). According to the report of UNAIDS(2010) 1.8 million People died of AIDS , which indicate a 21% decrease since 2005, and new HIV infections rate have declined by more than 20% since their peak in 1997, and declined by 15% between 2001 and 2010. Still, there were about 2.7 million new infections in 2010 or more than 7,000 new HIV infections per day (UNAIDS, 2011).

2.2.1. Global Response of HIV Epidemic

With this regard, fight towards HIV/AIDS pandemic must adopt an approach that emphasizes the collective responsibility of individuals, community groups, different levels of government and other agencies to diminish irredeemable diseases (GHIVR, 2011).

Region	Total No. (%) Living with HIV end of 2010	Newly Infected in 2010	Adult Prevalence Rate 2010
Global Total	34 million (100%)	2.7 million	0.80%
Sub-Saharan Africa	22.9 million (67%)	1.9 million	5.00%
South/South-East Asia	4.0 million (12%)	270,000	0.30%
Eastern Europe/Central Asia	1.5 million (4%)	160,000	0.90%
Latin America	1.5 million (4%)	100,000	0.40%
North America	1.3 million (4%)	58,000	0.60%
Western/Central Europe	840,000 (2%)	30,000	0.20%
East Asia	790,000 (2%)	88,000	0.10%
Middle East/North Africa	470,000 (1%)	59,000	0.20%
Caribbean	200,000 (0.6%)	12,000	0.90%
Oceania	54,000 (0.2%)	3,300	0.30%

Table 2. 1 HIV Prevalence & Incidence by Region (extracted from UNAIDSa ,2011)

As a result, from the beginning of the 21st century, the international community identified and declared HIV as the formidable health and development challenge. A rapidly expanding HIV epidemic was already dramatically reversing decades of progress on key development indicators, such as infant mortality and life expectancy (UNAIDSb, 2010).

Although the global incidence of HIV infection had peaked in the mid-1990s, more than 3 million people were being newly infected per year, AIDS had become one of the leading causes of adults dying in sub-Saharan Africa and the full assault of the epidemic would not be felt until 2006, when more than 2.2 million people died each year from AIDS-related causes (UNAIDSP,2011).

The revolution in HIV treatment brought about by combination of antiretroviral therapy in 1996 had forever altered the course of disease among those living with HIV in high-income countries but had only reached a fraction of people in low and middle-income countries, which bore 90% of the global HIV burden (UNAIDSb, 2010).

According to the report of 13th International AIDS Conference in July 2000 in Durban, South Africa (see details at www.nlm.nih.gov/news/aidsnote00.html), activists, community leaders, scientists and health care providers joined forces to demand access to treatment and an end to the enormous health inequities between the global North and global South. Months later, world leaders established the Millennium Development Goals, a series of ambitious, time-bound targets aimed at achieving progress on several health and development goals over the next 15 years, including Millennium Development Goal: combat HIV, malaria and other diseases (GHSSH, 2011).

In 2001, the United Nation General Assembly Special Session on HIV/AIDS approved commitment with common targets in specific technical areas, such as expanding access to antiretroviral therapy, antiretroviral prophylaxis to prevent the mother to- child-transmission of HIV and HIV prevention (Schneider et al., 2010).

The member of the general assembly declared a dedicated global health fund to finance the HIV response, resulting in the launch of the global fund to fight AIDS, tuberculosis and malaria one year later: The global fund quickly became a cornerstone in the global response to HIV, funding country-led responses through a pioneering, performance-based grant system.

According to Schneider et al. (2010) in 2003 the United States Government announced the United States President's Emergency Plan for AIDS relief. At US$ 15 billion over five years, it was the largest single funding commitment for a disease in history. The United States President's Emergency Plan for AIDS Relief was reauthorized in 2008 for up to US$ 48 billion to combat AIDS, TB and malaria for 2009–2013 and supports strategic interventions in the drugs and diagnostics markets in 94 countries. Besides, the global HIV/AIDS response indicates (GHIVR, 2011) that there is an immediate response politically as well as financially towards against HIV epidemic. The commitment is in parallel with normative guidance and strategic technical

innovations, including a ground-breaking approach to scaling up treatment access in low- and middle-income countries: the public health approach to antiretroviral therapy.

To mention the vital elements of the public health approach includes: using standardized treatment protocols and drug regimens; simplified clinical monitoring; maximizing coverage with limited resources; optimizing human resources for health and involving people living with and affected by HIV in designing and rolling out antiretroviral therapy programs (USAIDHPI, 2010).

As mentioned in the general fact of HIV/AIDS across the world, it is not the right time to keep silent about this incurable disease. However, everybody should do or react something against HIV/AIDS. For same reason, the researcher of this experimental study is striving from part towards the fight against HIV by implementing the possible and available Technology to the problem domain.

2.3. The Status of HIV/AIDS Epidemic in Ethiopia

In Ethiopia HIV/AIDS infections were first identified in 1984, and the first AIDS cases were reported in 1986 (Garbus, 2003). The prevalence of the disease was low in 1980s, but it escalated quickly through the 1990s. As a result, it rose from an estimated 3.2% of the adult population in 1993 to 7.3 % by the end of 1999(MOH, 2000).

According to NIC (2002) Ethiopia is classified (along with Nigeria, China, India and Russia) as belonging to the 'next wave countries' with large populations at risk from HIV infection which eclipse the current focal point of the epidemic in central and southern Africa. It is estimated that seven to 10 million Ethiopians be infected by 2005 because of the current high adult prevalence rate, widespread poverty and low educational levels (Garbus, 2003).

The dominant mode of transmission is through heterosexual contact (estimated to account for 87% of infections) and mother to child transmission (MTCT) (10% infections) (GoE, 2004). However, the prevalence of HIV/AIDS has reduced gradually. Ethiopian Ministry of Health estimates that the current adult HIV prevalence decline to 3.5% (MOH,2006). This figure jumps to an estimated 5% among pregnant women, but uptake of antiretroviral prophylaxis for Preventing Mother-to-child Transmission of HIV (PMTCT) has been minimal and the rate of HIV transmission to

children born to HIV positive women remains at 25 percent. According to the report of Ministry of health two million Ethiopian peoples are living with HIV-positive, but among this it is estimated that fewer than 10% know their HIV status (MOH, 2006).

The above report suggests that the percentage of prevalence rate increase in 1990s from three to seven, but Since 2005 HIV prevalence rate was reduced to 3.5 % (MOH, 2006). According to the current report of UNAIDS, Ethiopia has an estimated two million people living with HIV and the third highest number of infections in Africa. With a population of 83 million people and per capita income of less than US$100 annually, it is also one of the world's poorest countries (UNAIDSE, 2011).

2.3.1. Government Response Towards HIV/AIDS Epidemic in Ethiopia

The government's response to HIV/AIDS was immediate, but inadequate. A National Task Force was established in 1985 following the report of HIV prevalence in the country. Efforts were made, although on a limited scale, to expand information, education and communication (IEC), condom promotion, surveillance, patient care, and HIV screening laboratories at different health institutions(Hailom et al.,2005).

As part of these efforts to address the epidemic, the Government of Ethiopia has expanded HIV/AIDS Prevention and Control Offices at national, regional, zonal and sub-city/wereda levels. The overall objectives of the National office are to guide the implementation of the successful programs to prevent the spread of HIV/AIDS, decrease vulnerability of individuals and communities to the epidemic, provide care services for those living with the virus and reduce the adverse socio-economic consequences of the epidemic (NAC, 2001).

Besides that, the Ethiopian Ministry of Health has set a strategic plan for year 2007 targets, which include: encourage individuals to receive HIV counseling and testing; Address and create opportunity for those people who has getting HIV positive to be on antiretroviral treatment and give particular attention for HIV positive pregnant women to be treated with a complete course of antiretroviral prophylaxis to prevent transmission from mother to child (MOHHA, 2006).To achieve this goal, the Ministry of Health has an ambitious roadmap for scale-up of HIV prevention,

treatment and care that plans the delegation of responsibility for achievement of national targets from central hospitals outwards to the local health centers and private clinics.

Although there is a challenge to address evenly HIV prevention, testing and care in Ethiopia. Eighty-five percent of the population lives in rural areas and suffers from a severe lack of access to public health services. There is also a critical shortage of physicians (an estimated 1,200 in public service practice for a population of 83 million) and other trained health care workers. In addition to this, the per capita expenditures for health from all sources is only US$5.60 compared to US$12.00 per person in the Africa region as a whole (Hailom et al.,2005).

In response to the extreme shortage of qualified health care workers, the Ministry also enacted a plan to further decentralize its operations through collaborations with community-based NGOs. The Ministry of Health also encouraged African Services to replicate its highly successful prevention outreach and VCT model at additional sites and to add diagnosis and prophylaxis of opportunistic infections and antiretroviral therapy to its VCT services.

2.4. Voluntary HIV Counseling and Testing Service

A greater knowledge of HIV status within a community is critical to expanding access to HIV treatment, care and support in a timely manner as it offers people living with HIV an opportunity to receive information and tools to prevent HIV transmission to others [WHOa 2010].

The most successful method of gaining information about HIV has been through Voluntary Counselling and Testing (VCT). According to Mugula et al.(1995) Voluntary HIV counselling and testing is the process by which an individual undergoes counselling enabling him or her to make an informed choice about being tested for HIV. This decision must be entirely the choice of the individual and he or she must be assured that the process be confidential. The scope of VCT service includes :HIV pre-test counselling, testing, and post-test counseling. It is an entry point for prevention and care (FHI, 2004).

2.4.1. The process of VCT service

In HIV counseling, although the goal is eliminating risk, it is recognized that this can be best achieved through small steps for incremental behavior change that bring about risk reduction.

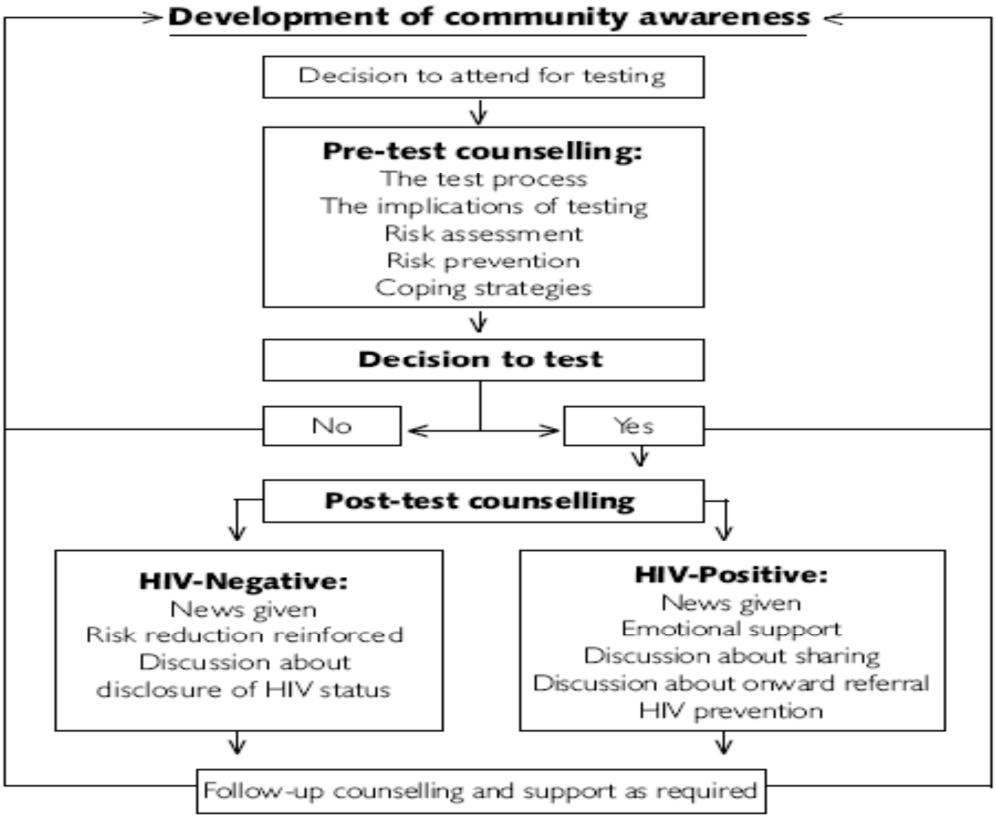


Figure 1. 1 Diagrammatic overview of the process of VCT Service

Source: (http://data.unaids.org/Publications/irc-pub01/jc379-vct_en.pdf)

As shown in the above diagram, the VCT process consists of pretest, post-test and follow-up counselling. HIV counselling can be adapted to the needs of the client/s and can be for individuals, couples, families and children and should be adapted to the needs and capacities of the settings in which it is to be delivered (UNAIDSHIV, 2000).

The content and method may vary considerably for men and women and with various groups, such as counselling for young people, men who have sex with men (MSM), injecting drug users (IDUs) or sex workers. The manner in which understanding of HIV sero status is very important in facilitating adjustment to impact of HIV infection.

Counselling as part of VCT ideally involves at least two major sessions (pretest counselling and post-test counselling). More time can be offered before or after the test, or during the period the client is coming up for test results.

Pre-test counselling: It should be offered before taking an HIV test. In an ideal world, the counsellor prepares the client for the test by explaining what an HIV test is, as well as by correcting myths and misinformation about HIV/AIDS. They may also discuss the client's personal risk profile, including discussions of sexuality, relationships, possible sex and/or drug-related behavior that increase risk of infection, and HIV prevention methods. The counsellor discusses the implications of knowing one's sero status, and ways to cope with that new information.

People who do not want pre-test counselling should not be prevented from taking a voluntary HIV test. However, informed consensus from the person being tested is usually a minimum ethical requirement before an HIV test.

Post-test counselling: The main goal of this counselling stage is very crucial to help clients understand their test results and initiate adaptation to their seropositive or negative status. It should always be offered.

When the test is seropositive, the counsellor tells the client the result clearly and sensitively, providing emotional support and discussing how he/she cope. During this session the counsellor must ensure that the person has immediate emotional support from a partner, relative or friend. When the client is ready, s/he may offer information on referral services that may help clients accept their HIV status and adopt a positive outlook.

In other way round, Counselling is also important when the test result is negative. While the client is likely to feel relief, the counsellor must emphasize several points. Here, discussion about how to changes in behavior that can help the client stay HIV-negative, such as safer sex practices including condom use and other methods of risk reduction. S/he also motivates the client to adopt and sustain new, safer practices and provide encouragement for these behavior changes.

During the “window period” (approximately 4-6 weeks immediately after a person is infected), antibodies to HIV are not always detectable. Thus, a negative result received during this time may not mean the client is definitely uninfected, and the client should consider taking the test again in 1-3 months.

2.4.2. Benefit of VCT Service

According to Mugula et al. (1995) VCT has a vital role to play within a comprehensive range of measures for HIV/AIDS prevention and support, and should be encouraged. It is an efficient and cost-effective strategy in expanding access to prevention, treatment and care services.

The potential benefits of voluntary counselling and testing service (FHI, 2004; Mugula et al., 1995) includes: facilitates behavior change, reduces the stigma attached to those who live with HIV/AIDS (Kathy and Stacey, 2009) and early VCT can lead to a delay in HIV deaths, better ability to cope with HIV related anxiety; awareness of safer options for reproduction and infant feeding; and motivation to initiate or maintain safer sexual and drug related behaviors. Other benefits include safer blood donation

World Health Organization (WHOS, 2005) summarized the benefits of knowing HIV status in three different levels such as at the individual, community and population.

- **At the individual**: VCT enhanced ability to reduce the risk of acquiring or transmitting HIV; access to HIV care, treatment and support; and protection of unborn infants.
- **At the community**: it creates a wider knowledge of HIV status and its links to interventions can lead to a reduction in denial, stigma and discrimination and to collective responsibility and action.
- **At the population level**: knowledge of HIV epidemiological trends can influence the policy environment, normalize HIV/AIDS and reduce stigma and discrimination.

2.5. The prevalence of HIV/AIDS in Ethiopian Higher education

Study show that the highest concentration of HIV positive people is between the ages of 15-40 (MOH, 2000). Besides, scholars in many part of the world point out that because of the high

concentration of people in this economically significant segment of the population of labor force age, productivity has been and continues to be negatively impacted by the factor such as HIV-related absenteeism, disability and the early death of experienced workers (Saint,2005).

Producing high level human resource to run an entire economy has relied on higher education. As a result society is undertaking and performing task effectively and efficiently based on the task of the organization missions.

Furthermore, HIV/AIDS holds the potential to undermine the country's substantial investments in education. It affects all people all over the world without discrepancy in race, religion and gender. Overall, it has major impact on distribution of education.

AIDS now exists within all regions of Ethiopia. The estimated national HIV infection rate is 2.1% (UNAIDS, 2007). This rate is roughly 1 in 50 Ethiopian adults aged range from 15 to 30 has the HIV virus. The highest at risk for HIV infection are female sex workers, children born from HIV-infected parents, and injection drug users.

Higher education communities are particularly vulnerable to HIV/AIDS due to their age groups. This vulnerability introduces a sizeable risk to the expected returns on investments made by families and government in the education of university or college students (Otaala, 2003). In spite of this risk, universities in Ethiopia have not established institutional policies or programs for the management and prevention of HIV/AIDS until 2005 (Saint, 2005). Institutional policies for the management of HIV/AIDS plays significant role for reduction of HIV infection.

This policy may include review of regulations on sick leave, confidentiality and the rights of persons living with AIDS, student counseling services to awareness programs, and curriculum content on testing facilities (Otaala, 2003).

According to Saint (2005) little has been known about the status of AIDS on Ethiopian University. As indicated by recent study in Jimma university which estimated that 12.2 % of the university communities were HIV positive (MOE, 2004). Since then Ministry of Education and university officials are discussed on issue to react against HIV infection in higher education.

As the result of this Addis Ababa university (AAU) became the pioneer that started the first HIV/AIDS policy in Ethiopia higher education at 2006 (AAU,2006). The policy document is established with a collaboration of African AIDS Initiatives International (AAII), but it started its VCT service in AAU in 2005.

According to the report of Addis Ababa highlight (AAH, 2006), the HIV/AIDS policy document served as a base and policy for other institutions in Ethiopia and out of Ethiopia, in the hope that other institutions of higher learning would follow in the footsteps of Addis Ababa University for the benefit of their young people.

As the oldest institution of higher learning in Ethiopia, Addis Ababa University played the leading role in promoting instruction, training and research in the fight against HIV epidemic.

The current student population of Addis Ababa University enrolled in the graduate, undergraduate, extension, and summer programs adds up to over 50,000. At present, AAU has nearly 2000 full time academic staff and over 5000 administrative personnel (AAUa, 2011).

The university is committed to playing active role in the fight against HIV/AIDS primarily by engaging in research in biomedical and psychosocial spheres. In doing this, it needs to establish collaborative relationships with scientific institutions both nationally and abroad. It also needs to work closely with national, regional and international partners in mitigating the impact of HIV/AIDS on its own staff and students as well as the wider Ethiopian society.

In this regards, the university encourage faculty, staff and students to submit to voluntary counseling and testing (VCT) and facilitate this by establishing the necessary facilities on campus. After establishment of a formal HIV policy document, various non-governmental organizations has involved in ant-AIDS movement. In this respect, an African AIDS initiatives International take the line share against mitigating HIV infection in higher education particularly in AAU.

As introduced in chapter one background of the study 1.1, AAII provides HIV prevention related service for the university community. According to Data Manager of AAII, the organization has two primary VCT centers in university i.e. 6 kilo and 4 kilos. Besides, it provides a VCT service outside the university community.

The organization collects and organizes client's record provided in each site on daily base. As the researcher observed the database of Epi Info software of the organization holds an estimated of 23,000 clients are visited the AAII VCT center both in campus and off campus site.

The AAII VCT center provides six days per week and 24 days per month services for university community. Its major activities include provision of quality VCT service; implementing quality information management system; referral network system establishment and monitoring feedback; strong VCT HIV sensitization and promotion; and replication of best practices to other centers.

Currently, the organization extends its VCT service from AAU to other Ethiopian higher education including Dbere Marko's University and Mezan Tepi University.

Africa AIDS Initiative International Inc.

TOTAL NUMBER OF VCT CLIENTS COUNSELED AND TESTED

FROM MARCH 2005 - OCTOBER 31/2011

	Number of Clients Counseled			Number of Clients Counseled & Tested			Test Result						Post test Counseled		
							Sero-Positive			Sero-Negative					
Category	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T
Student	9652	4433	14085	9554	4402	13956	43	41	84	9511	4361	13872	9554	4402	13956
Staff	1078	816	1894	1073	809	1882	26	29	55	1047	780	1827	1073	809	1882
Community	4436	2771	7207	4404	2759	7163	199	266	465	4205	2493	6698	4404	2759	7163
Total	15166	8020	23186	15031	7970	23001	268	336	604	14763	7634	22397	15031	7970	23001

Table 2. 1 Annual report of AAII VCT Client from March 2005-October2011

The above table indicates that the number of student who has HIV positive test result is greater than from that of staff but less than community. Besides, the total number of client who checks their status in this range is better than other user of the center. According to this report the percentage of females (56%) are more exposed for HIV than male (44%).

The data in all centers are collected using the same standard paper format that are used by CDC (Center of Disease control and Prevention) and managed by VCT EPi Info software system.

As shown in table 2.2 the initial data are collected from AAII VCT centers for seven years. The total clients visited the centers up to October, 2011 are 23,000.

The data collected over time from client can tell so much if appropriately analyzed as to what factor predominantly cause HIV infection. However, when data is getting larger and larger, the data carry a lot of implicit knowledge that cannot be accessed with simple statistical analysis methods. Hence application of other tool is inevitable. So the most appropriate technique to discover and manage knowledge implicitly hidden inside the collected data is data mining (Abdule-kereem et al.,2000).

We discussed the importance of VCT service in section 2.4.2, the data that we obtain from this center is very crucial and play significant role in decision making related to HIV/AIDS control and prevention.

CHAPTER THREE

LITERATURE REVIEW

In this chapter, an attempt has been made to review the literature on the concepts and techniques of data mining in general and its application in the health care sector in particular which is aimed to provide background about the models to be built.

3.1. Overview of Data Mining

Across a wide variety of fields, data were being collected and accumulated at a dramatic pace. The rapid growth and integration of databases provides scientists, engineers , and business people with a vast new resources that can be analyzed to make scientific discoveries, optimize industrial systems and uncover valuable patterns **(**Hand et al ., 2001).

The developments in digital data acquisition and storage technology have resulted in the growth of huge databases. This has occurred in all area of human attempt , from the ordinary (such as supermarket transaction data, credit card usage records, telephone call detail, and government statistics) to the more exotic (such as images of astronomic bodies , molecular databases, and medical records)(Thearling et al. ,nd).The large size and complexity of such data in many scientific domains becomes impractical to manually analyze, explore, and understand data.

To undertake large data analysis projects, researchers and practitioners have adopted established algorithms from statistics , machine learning, neural networks ,and databases and have also developed new methods targeted at large data mining problems **(**Hand et al., 2001). Witten and Frank (2005)states that lack of data is no longer a problem at a current stage. However, the inability to generate useful information from data is the problem. On other hand, it was recognized that information is at the heart of business operation and that decision make use of the data stored to gain valuable insight in to the business (Rea, 1995). Thus there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data (Fayyad et al., 1996). These theories and tools are the subject of the emerging fields of knowledge discovery in database and data mining.

3.2. Knowledge Discovery in Database (KDD) and Data mining

Conventionally the idea of searching pertinent patterns in data has been referred using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely (Fayyad et al, 1996).

The term knowledge discovery in databases (KDD) is synonymously used with data mining. The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro, 2000). The purpose was to emphasize that knowledge is the end product of a data-driven discovery and it has been popularized in the 'artificial intelligence' and 'machine-learning' fields.

In Fayyad's et.al. (1996) view, KDD refers to the overall process of discovering useful knowledge from data and that data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

Trybula (1997) also states that knowledge discovery (KD) is the process of transforming data into previously unknown or unsuspected relationships that can be employed as predictors of future action. Furthermore, he asserts that KDD is a term that has been employed to encompass both data mining and KD. Essentially, the basic tasks of data mining and KD are to extract particular information from existing databases and convert it into understandable or sensible conclusions (i.e., knowledge).

3.3. What is Data mining?

In the Information Technology era information plays vital role in every sphere of the human aspects. Thus, to efficiently inspire information, it is very important to generate information from massive collection of data. The data can range from simple numerical figures and text documents to more complex information such as spatial data, multimedia data, and hypertext documents (Deshpande and Thakare, 2010). However, the huge size of these data sources makes it impossible for a human experts to come up with interesting information or patterns that help in the proactive decision making process. Therefore, to take complete advantage of data; the data retrieval is not enough. It requires a tool for automatic summarization of data, extraction of the essence of

information stored, and the discovery of patterns in raw data. This tool is called data mining (Deshpande and Thakare, 2010).

Data mining is one step in the KDD process (Fayyad et al., 1996). It is the most researched part of the process. Data mining can be used to verify a hypothesis or use a discovery method to find new patterns in the data that can do predictions with new, unseen data.

Many definitions of data mining could be found in the literature, Berry et.al. (1997) define it as the exploration and analysis of large quantities of data by automatic or semiautomatic means in order to discover meaningful patterns and rules. According to Bigus (1996), data mining is the efficient discovery of valuable, non-obvious information from a large collection of data.

Witten (2000) also defines as data mining is the analysis of large observational data sets with the goal of finding unsuspected relationships. A data set can be "large" either in the sense that it contains a large number of records or that a large number of variables is measured on each record.

Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data; The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases (Two Crows Corp ,1999); It is the process of finding mathematical patterns from (usually) large sets of data; these can be rules, affinities, correlations, trends, or prediction models (Berry and Linoff,2000).

Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (WKDD ,2008).

3. 4. Knowledge Discovery Process Models (KDPM)

The knowledge discovery process can be formalized into a common framework using process models. The main reason for introducing process models is to formalize knowledge discovery projects within a common framework, a goal that result in cost and time savings, and improve understanding, success rates, and acceptance of such projects. The models emphasize independence from specific applications, tools, and vendors (Cios et al., 2007).

A KD process model provides an overview of the life cycle of a data mining project, containing the corresponding phases of a project, their respective tasks, and relationships between these tasks. There are different process models originating either from academic or industrial environments. All the processes have in common the fact that they follow a sequence of steps which more or less resemble between models

Although the steps are executed in a predefined order and depend on one another (in the sense that the outcome of a step can be the input of the next step), it is possible that, after execution of a step, some previous steps are re-executed taking into consideration the knowledge gained in the current step (feedback loops).

An interesting issue that arises in any KDD process model is choosing which algorithm or method should be used in each phase of the process to get the best results from a dataset. Since taking the best decision depends on the properties of the dataset, Popa et al. (2007) proposes a multi-agent architecture for recommendation and execution of knowledge discovery processes.

Knowledge discovery in database (KDD) and data mining is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the internet and the widespread use of databases have created an immense need for KDD methodologies. In this paper, we provide an overview of common knowledge discovery process model and approaches to solve tasks. The challenge of extracting knowledge from data draws upon research in statistics and database (Mamcenko and Beleviciute, 2007).

A as result, data mining research require stable and well defined foundation which are well understood and popularize thought the community (Kurgan &Musilek,2006). The primary objective of data mining process or methodologies is building stable models following some logical steps (Berry &linoff,2004) .

According to Santos & Azevedo (2008) the followings are commonly used Data mining process model / methodology /KDPM that data mining appliedin research; Knowledge Discovery in Database (KDD), CRoss-Industry Standard Process for Data Mining(CRISP-DM), Sample, Explore, Modify, Model, Assess(SEMMA) and Hybrid model.

3.4.1. KDD Process Model

The essence of Academic research model was started in the mid-1990s, when the DM field was being shaped; researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain (Cios et al., 2007). As a result different Academic research models are developed and followed by industry. According to Fayyad et al. (1996). KDD process model was the leading Academic research model applied in different data mining project.

KDD process, as presented in (Fayyad et al, 1996), is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are five stages as presented in figure 3.1:

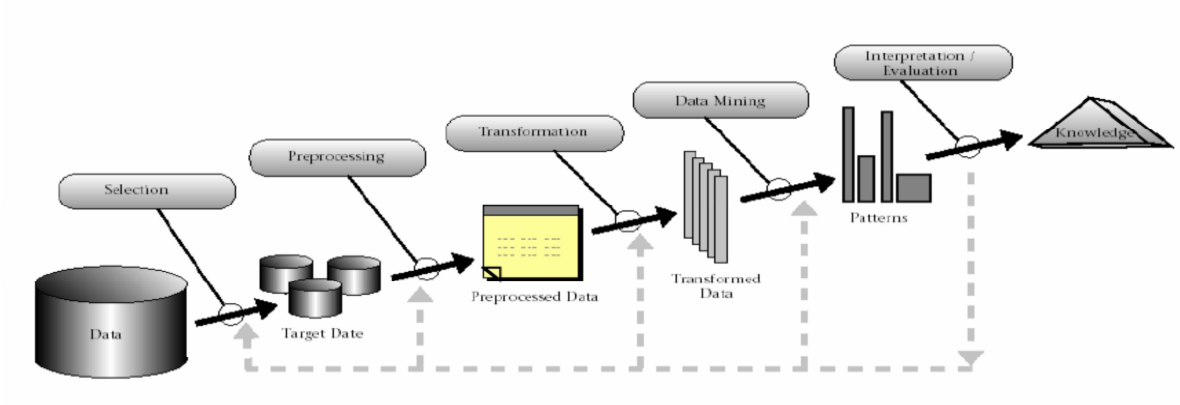


Figure 3. 1 Overview of KDD

An outline of the steps in Figure 3.1 has been adequate for understanding the concepts required for the KDD process. The following are the steps involved:

Selection - this stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed;

Pre-processing - this stage consists on the target data cleaning and preprocessing in order to obtain consistent data. Here we also try to eliminate noise that is present in the data. Noise can be defined as some form of error within the data. Some of the tools used here can be used for filling missing values and elimination of duplicates in the database.

Transformation - this stage consists on the transformation of the data using dimensionality reduction or transformation methods. Usually there are cases where there are a high number of attributes in the database for a particular case. With the reduction of dimensionality we increase the efficiency of the data-mining step with respect to the accuracy and time utilization.

Data Mining - The data mining step is the major stage in data KDD. This is when the cleaned and preprocessed data is sent into the intelligent algorithms for classification, clustering, similarity search within the data, and so on. Here we chose the algorithms that are suitable for discovering patterns in the data. Some of the algorithms provide better accuracy in terms of knowledge discovery than others. Thus selecting the right algorithms can be crucial at this point.

Interpretation/Evaluation –in this stage the mined data is presented to the end user in a Human-viewable format. This involves data visualization, which the user interprets and understands the discovered knowledge obtained by the algorithms.

3.4.2. CRISP-DM

A second most popular process model that has proven application is Cross Industry Standard Process Data Mining (CRISP-DM). CRISP-DM is a product neutral data mining model developed by consortium of several companies (Roiger and Geatz, 2003). It is a standard process model in industries which consisting of a sequence of steps that are usually involved in a data mining study .CRISP-DM was developed by the means of the efforts of an association initially composed with Daimler Chrysler, SPSS and NCR (Chapman et al.,2000). The life cycle of a data mining project consists of six phases. These are business understanding, data understanding, data preparation, modelling, evaluation and deployment .CRISP-DM process consists of six phases (Roiger and Geatz, 2003).

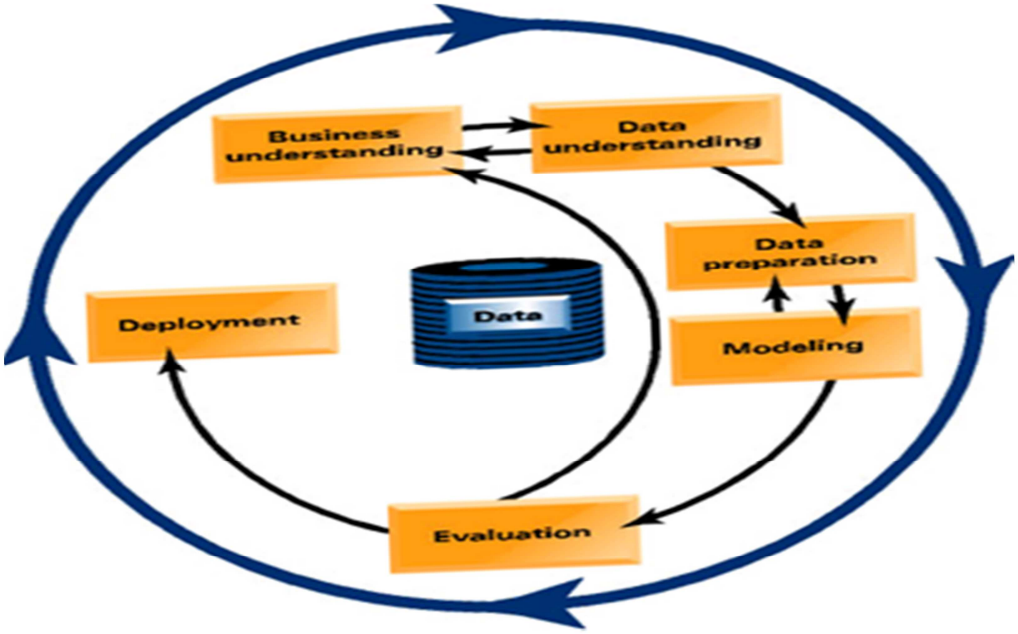


Figure 3. 2 CRISP-DM process model adapted from (Daniele and Satheesh ,n.d.)

CRISP-DM is complete and well documented. All this stages are duly organized, structured and defined, allowing that a project could be easily understood or revised (Santos & Azevedo, 2005).

As seen from the figure 3.2, CRISP-DM model is divided into six phases and related tasks as described in the [Daniele and Satheesh ,(n.d.); Chapman et al.,2000] as follows:

1. **Business understanding**: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives [Daniele and Satheesh ,(n.d.)]. Some of the objectives of this phase are determine business objectives and business success criteria assess situation such as inventory resource requirements, assumptions constraints risk contingencies cost and benefits determine data mining goals produce project plan initial assessment of tools and techniques. Business understanding is critical because it identifies the business objectives and, thus, the success criteria of data mining projects (Koh, and Tan, 2005).

2. Data understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. The major activities are collect initial data (initial data collection report), describe data (data description report), explore data (data exploration report) and verify data quality (data quality report).

3. Data preparation: The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order. Data preparation and cleaning is an often neglected but extremely important step in the data mining process models. Often, the method by which the data where gathered was not tightly controlled, and so the data may contain outlier values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes). Analyzing data that has not been carefully screened for such problems can produce highly misleading results, in particular in predictive data mining [Daniele M. and Satheesh R,.nd)].

The main tasks in data preparation are; select data which is rationale for industry exclusion or inclusion clean the data and document it, construct data derived attributes and generated reports integrate and merge data, format data reformat data and have dataset description. Data preparation step is by far the most time-consuming part of the knowledge discovery process (Cios et al., 2007).

4. Building Model: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phases such as select modeling techniques and modeling assumptions, generate test design, build model, parameter settings and model description and assess model, model assessment, revise parameter settings.

5. Evaluation: Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives. At the end of this phase, a decision should be reached on how to use the DM results through assessment of data mining results with respect to business success criteria to

approve models, review of process, and determine next steps such as possible actions and decisions making based on patterns and relationships.

6. Deployment: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained need to be organized and presented in a way that the end user can use and understand it. The main activities in this phase are plan deployment, plan monitoring and maintenance (monitoring and maintenance plan), produce final plan and present, review project and produce documentation. According to Chapman et al. (2000), depending on the requirement, these phase can be as simple as generating a report as complex as implementing a repeatable data mining process across the enterprise.

According to kurgan and Musilek (2006) the CRISP-DM methodology is the most suitable process model for novice data miners due to its easy, complete documentation , and intuitive , industry-applications-focused description.

The CRISP-DM model has been used in domains such as medicine, engineering, marketing, and sales. It has also been incorporated into a commercial knowledge discovery system called Clementine (see SPSS Inc. at http://www.spss.com/clement and Cios et al., 2007).

3.4.3. The SEMMA Process

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA stands for(**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess) refers to a methodology that clarifies this process (Obenshain, 2004). The SAS Institute considers a cycle with five stages for the process:

Sample – this stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.

Explore - this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify - this stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model - this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals (Santos & Azevedo, 2005).

3.4.4 Hybrid model

The development of both academic particularly the KDD and industrial oriented (CRISP-DM and other) data mining models has led to the growth of hybrid models, i.e., models that combine the features and job of both. Hybrid model is a six-step KDP model (see Figure 3.3) developed by Cios et al.(2007). It was developed mainly based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include introducing several new explicit feedback mechanisms and in last steps the knowledge discovered for a particular domain may be applied in other domains.

According to Cios et al.(2007) the description of the six steps are explained below.

Step 1: understanding of the problem domain. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. The major tasks undertaken in this phase are: Understand and learn domain-specific terminology.

- explain the problem area in detail and make a boundary for it
- Lastly, convert the project goals into DM goals, and select appropriate data mining tools to be used in performing the whole research.

Step 2: understanding of the data. In this phase collecting sample data and deciding the types, format and size of data was a primary task. Background knowledge can be used to guide these efforts. Besides, data should be checked for completeness, redundancy, missing values,

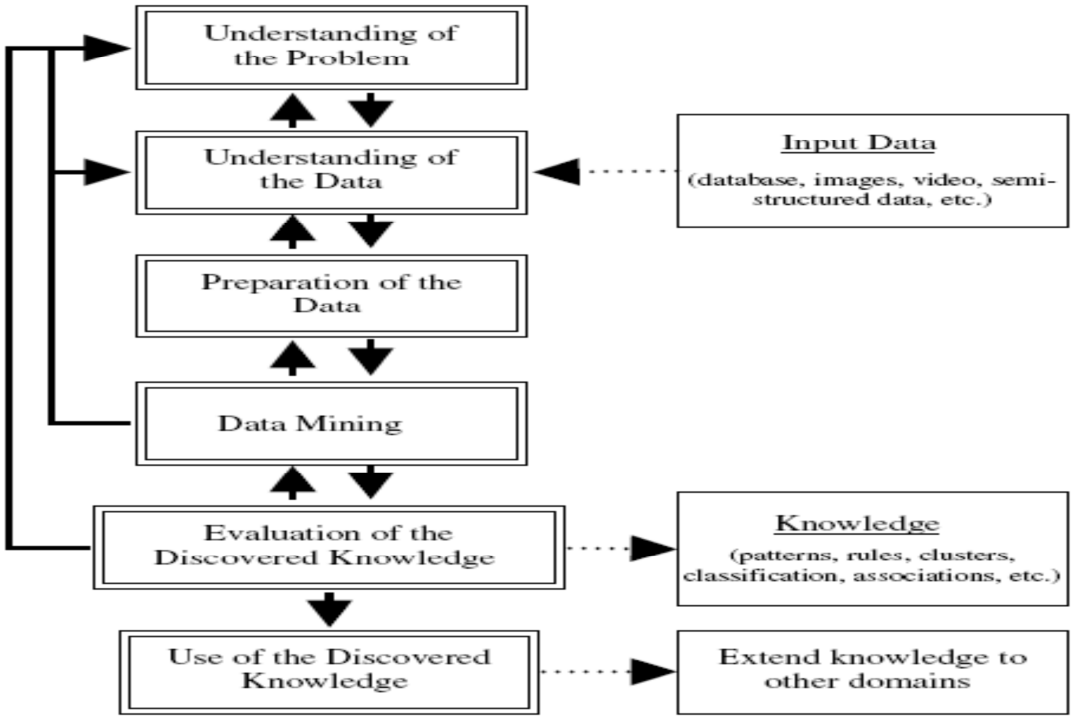


Figure 3. 3 The six phases of Hybrid data mining model(Pal, et al.,2005)

plausibility of attribute values, etc. verification of the usefulness of the data with respect to the DM goals was also a major activity undertaken in this step.

Step3: Preparation of the data. Once the problem is understand and selected complete and appropriate data to the problem under investigation the next step be selecting representative data. As a result, the researcher be used only the selected data as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization).

The end results are data that meet the specific input requirements for the DM tools selected in Step 1.This steps is the most time consuming part of all data mining model (Cios et al., 2007).

Step4 Data mining: Another key step in the knowledge discovery process is data mining. It involves the use of several DM tools on data prepared in step 3.The application of this tools discover new knowledge. The process of discovering new information includes: the data model was constructed using one of the chosen DM tools and training and testing procedures are designed. Then generated data model was verified by using testing procedures. It takes less time than data preprocessing steps. Some of the common data mining technique implemented using data mining tools are decision tree, neural networks, clustering, support vector machine, rule induction, association rule.

Step 5 Evaluation of the discovered knowledge. Evaluation includes understanding the results, checking whether the discovered knowledge was novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process was revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is also prepared.

Step 6 Use of the discovered knowledge. The final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge was created and the entire project will be documented and reported.

3.4.5. Comparison of SEMMA, KDD and CRISP-DM

Today, research efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made. The subsequent models provide more generic and appropriate descriptions. Most of them are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application (kurgan and Musilek ,2006). Santos & Azevedo (2008) have summarized the association of the three most popular process model steps as shown in Table 2.1

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

Table 3. 1 Comparison of Data mining process model

As shown in table 3.1 most of the steps to be followed in all methods look like similar. Nevertheless, KDD and SEMMA do not have a step "understanding the problem" before selection and sample steps respectively. Undoubtedly, understanding the domain problem is often a prerequisite steps but missed by the three methodology. Considering the presented analysis we conclude that SEMMA and CRISP-DM can be viewed as an implementation of the KDD process described by (Fayyad et al, 1996).

In general, According to Refaat (2007) all methodologies contain and followed same data mining tasks. Hence this research be conducted based on understanding of the domain problem and business context, consequently, the best fit data mining model to conduct the research is inevitably hybrid model. This model completes the above missed limitation in methodology.

3.5. Data Mining Task

The goal of data mining is to **learn** from data, and there are two broad categories of data mining strategies: supervised and unsupervised learning. Supervised learning methods are deployed (or strategically put into service in an information technology context) when values of variables (inputs) are used to make predictions about another variable (target) with known values. Unsupervised learning methods can be used in similar situations, but are more frequently deployed on data for which a target with known values does not exist (Obenshain, 2004).

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and

predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions (Han and Kamber, 2006; Kantardzic, 2003). According to Fayyad et.al. (1996), the data mining stage of KDD process consists of two main types of data mining methods: verification oriented (the system verifies user's hypothesis) and discovery oriented (the system finds new rules and patterns autonomously). The discovery-oriented methods can be further partitioned into descriptive (e.g. Visualization) and predictive (like regression and classification).

In predictive modeling tasks, one identifies patterns found in the data to predict future values (Levin and Zahavi, 1999). Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. Whereas descriptive techniques are sometimes referred to as unsupervised learning since there is no already-known result to guide the algorithms (Two crows corporation, 1999).

On the other hand, descriptive models belong to the realm of unsupervised learning. Such models interrogate the database to identify patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, among others, belong to this family of descriptive models (Levin and Zahavi, 1999; Han and Kamber, 2001).

The Predictive modeling is based on the use of the historical data. Predictive model data mining tasks include classification, regression, time series, and prediction. Unlike the predictive modeling approach, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties (Dunham, 2003). Clustering, summarization, association rules, and sequence discovery are viewed as descriptive in nature (shown in Figure 3.3: Dunham, 2003).

3.5.1. Classification and Prediction

Classification is one of the most common data mining tasks, which is also pervasive in human life. Classification finds common properties among different entities in order to organize them into predefined classes (Han et al. ,2001). Human beings usually classify or categorize in order to understand and communicate about the world. For any object or instance, classes are predefined according to the value of a specific field (Berry and Linoff, 1997).

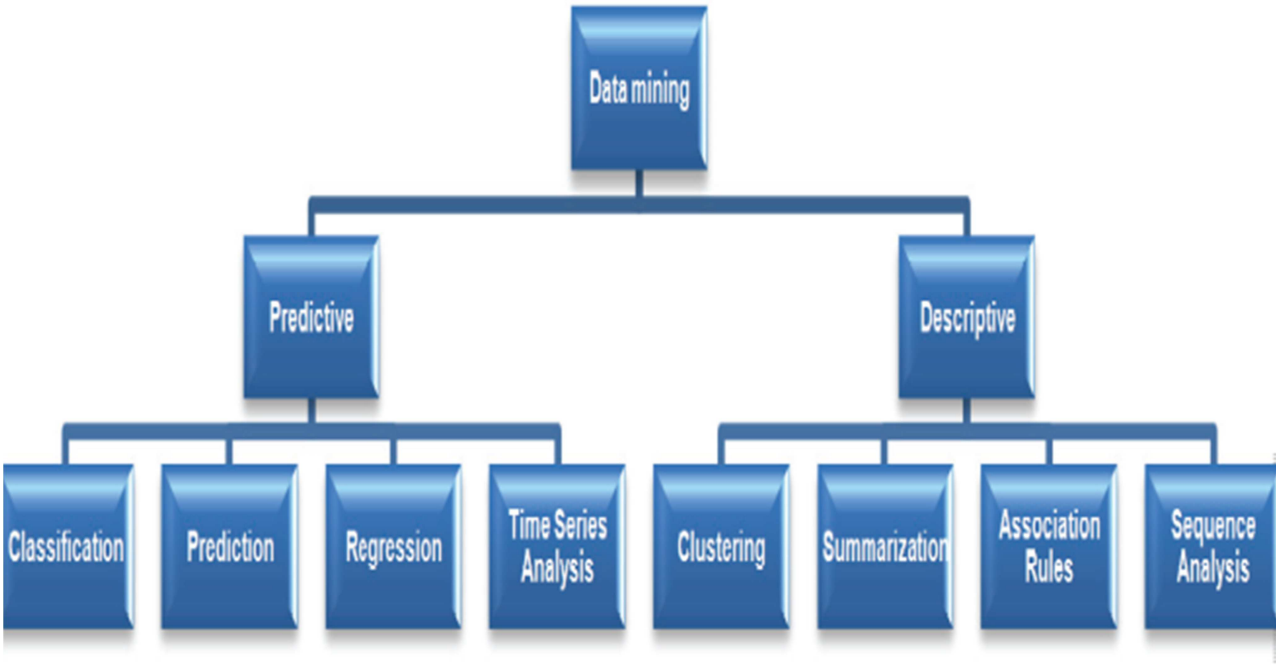


Figure 3. 4 Classification of Data mining models and tasks (Dunham, 2003)

There are many different methods, which may be used to predict the appropriate class for the objects or stations. Among the most popular ones are: logistic regression, decision tree, rule induction, case-based reasoning, and neural network (Moshkovich et al., 2002).

Even though , there are many data mining tasks available and commonly used in various application, the researcher used decision tree and rule induction techniques in this research because their result is simple to explain for end user and these technique support the selected data mining tasks for this research. The most widely used techniques for classification are decision trees, neural networks and memory-based reasoning (Berry et al., 1997).

3.5.1.1. Decision Trees

One of the most commonly used data mining techniques for classification tasks are decision trees. It is a simple knowledge representation technique and classifies instances in to a finite number of classes (Larvac, 1998). In decision tree induction, the nodes of the tree are labeled with attribute names, the edges of the tree are labeled with possible values for the attributes and the leaves of the tree generate decision tree from a given set of attribute-value tuples. A decision tree is constructed

by repeatedly causing a tree construction algorithm in each generated node of the tree (Larvac, 1998 ; Two crows corporation,1999).

The classification is performed separately for each leaf, which represents a conjunction of attribute valued in a rule (Last and Kandel, 2002). Structurally, decision trees consist of two types of nodes; non-terminal (intermediate) and terminal (leaf). The former correspond to questions asked about the characteristic features of the diagnosed case. Terminal nodes, on the other hand generate a decision (Rudolfer et al., 2002). Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that be used to make predictions.

An important quality of decision trees is that they handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets (Two crows corporation, 1999). Rogers and Joyner (2002) also stated "tree based models are good at selecting important variables, and therefore work well when many of the predictors are irrelevant."

A common criticism of decision trees is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In addition, all splits are made sequentially, so each split is dependent on its predecessor (Two crows corporation, 1999).

3.5.1.1.1. Decision tree Construction Methods

All decision tree construction methods are based on the principles of recursively partitioning the dataset until homogeneity is achieved. The construction of decision tree involves the following three phases (Sheikh, 2008):

1) Construction phases: the initial decision tree is constructed in these phases, based on the entire training dataset. It require recursively partitioning the training set in to two or more sub- partitions using a splitting criterion, until a stopping criteria is met. The basic strategy of construction phase is described as follows. The tree starts as a single node representing the training sample. If the samples are of the same class, then the node become a leaf and is labeled with that class. Otherwise, the algorithms uses entropy-based measure known as information

gains as a heuristic for selecting the attribute that best separate the sample s into individuals classes. The information gain measure is used to select the test attribute at each node in the tree. The attribute with the highest information gain is chosen as the test attribute for the current node.

2) Pruning phase: The pruning phase involves removing some of the lower branches and nodes to improve performance.

3) Processing the pruning tree: In this stage decision tree is proposed to improve understandability. Various listed decision tree algorithms such as CHAID, C4.5/C5.0, CART, ID3, J48 and many other with less familiar algorithms produce tree that differ from one another in the number of splits allowed at each level of tree , how those splits are chosen when the tree is built(Thearling et al. , n.d).

3.5.1.1.2. Decision tree Algorithms

The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items (Rokach and Maimon, 2005). A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), Quest, and C5.0,J48 (Two crows corporation, 1999).

J48 decision tree algorithms adopt an approach in which decision tree models are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built (Han & Kamber, 2006). According to Bharti et al. (2010), J48 decision tree algorithm is a predictive machine learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. It can be applied on discrete data, continuous or categorical data. This algorithms was used in this study to classify the VCT clients as HIV infected or non-HIV infected. The J48 decision tree can serve as a model for classification as it generates simpler rules and remove irrelevant attributes at a stage prior to tree induction. In

several cases, it was seen that j48 decision trees had a higher accuracy than other algorithms (Witten and Frank 2000). It offers also a fast and powerful way to express structures in data

To view the internal operation of the J48 algorithms Han and kamber (2006) explained how the algorithms works in detail step by step as follows.

Figure 3. 5 Basic pseudo code for J48 decision tree algorithms

Algorithm: Generate decision tree.

Input: Sets of training dataset (D), Attribute list ,Attribute selection method;

Output: A decision tree.

Procedures:(1) Create a node N; (2) If tuples in D are all of the same class, C then

(3) Return N as a leaf node labeled with the class C;

(4) If attribute list is empty then

(5) Return N as a leaf node labeled with the majority class in D; // majority voting

(6) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion;

(7) Label node N with splitting criterion;

(8) If splitting attribute is discrete-valued and multi way splits allowed then // not restricted to binary trees

(9) Attribute list ← attribute list- splitting_ attribute; // remove splitting attribute

(10) For each outcome j of splitting criterion // partition the tuples and grow subtrees for each partition

(11) Let Dj be the set of data tuples in D satisfying outcome j; // a partition

(12) If Dj is empty then (13) attach a leaf labeled with the majority class in D to node N;

(14) Else attach the node returned by Generate decision tree (Dj, attribute list) to node N; Endfor
(15) Return N;

In general, according to Han and Kamber (2006) the J48 decision tree algorithms follows five main procedures including:

1. Choose an attribute that best differentiates the output attribute values.

2. Create a separate tree branch for each value of the chosen attribute.

3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.

4. For each subgroup, terminate the attribute selection process if:

a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.

b. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.

5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

The algorithm is applied to the training data. The created decision tree is tested on a test dataset, if available. If test data is not available, J48 performs a cross-validation using the training data itself.

3.5.1.2. Neural Networks

A Neural Network is an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning (Rea, 2001). Neural networks are developed due to difficulty in creating the basic intelligence in computers using the conventional algorithmic approach (Pudi,2001). According to Berry and Linoff (1997), Bigus (1996) and Atiezenza et. al. (2002) Neural Networks are also probably the most common data mining classification.

Neural networks (NN) are those systems modeled based on the working of the human brain. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it.

The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input (Gupta et al., 2011).

These networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

3.5.1.3. Rule Induction

Rule induction is one of the techniques most used to extract knowledge from data, since the representation of knowledge as if/then rules is very intuitive and easily understandable by problem-domain experts (Wong and Leung, 2000). It is an area of machine learning in which formal rules are extracted from a set of observations. The extracted rules may represent a full model of the data (in the form of a rule set) like in Clark and Niblett (1989), or represent local patterns in the data (in the form of individual rules) like in Agrawal et al. (1996). Hence, regularities hidden in data are frequently expressed in terms of rules , rule induction is one of the fundamental tools of data mining at the same time (Grzymala-busse, n.d.). Usually rules are expressions of the form

if (attribute − 1; value − 1) and (attribute − 2; value − 2) and ……and (attribute − n; value − n) then (decision; value):

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain(Grzymala-busse, nd.). Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labeled as attributes and a decision. We restrict our attention to rule induction which belongs to supervised learning: all cases are pre classified by an expert (Grzymala-busse, n.d.). In other words, the decision value is assigned by an expert to each case. Attributes are independent variables and the decision is a dependent variable. Rule induction sometimes called rule learner is also perhaps the form of data mining that most closely resembles the process that most people think about when they think about data mining, namely “mining” for gold through a vast database. The gold in this case refers to the

rule that is interesting- that tell you something about your database that you did not already know and probably weren't able to explicitly articulate(Thearling et al. ,n.d).

Rule induction on a database is a process undertaking using intelligent software where all possible patterns are systematically pulled out of the data . In this process the accuracy is added to them that tell the user how strong the pattern is and how likely it is to occur again (Thearling et al. ,n.d). It has been widely used to represent knowledge in expert system and they have the advantage of being easily interpreted by human experts because of their modularity(Cabena etal.,1998; Moshkovich et al, 2002). Rule induction systems are highly automated and are probably the best of data mining techniques for exposing all possible predictive patterns in a database. They can be used in prediction problem but algorithm for combining evidence from a variety of rules come from practical experience (Thearling et al. ,n.d)

In rule induction, rules are expressed in the form of "if this, this and this then this". To make the rules useful, we require two important information including **accuracy** and **coverage.** This because the pattern in the database is expressed as rule does not mean that it is true all the time. Similar to other data mining algorithms it is important to recognize and make explicit the uncertainty in the rule, i.e. how often is the rule correct. This is what the "accuracy" of the rule means. The coverage of the rule has to do with how much of the database the rule "cover" or how often the rule applies (Thearling et al. ,n.d). Summary of coverage and accuracy be showed below.

	Accuracy Low	Accuracy High
Coverage High	Rule is rarely correct but can be used often.	Rule is often correct and can be used often.
Coverage Low	Rule is rarely correct and can only be used rarely.	Rule is often correct but can only be used rarely.

Table 3. 2 Rule coverage versus accuracy adapted from (Thearling et al. ,nd).

3.5.1.3.1. Rule induction for Prediction

The objective of data mining is to extract valuable information from one's data, to discover the 'hidden gold'. In Decision Support Management terminology, data mining can be defined as 'a decision support process in which one search for patterns of information in data' (Parsaye, 1997).

Data mining techniques are based on data retention and data distillation. Rule induction models belong to the logical, pattern distillation based approaches of data mining. These technologies extract patterns from data set and use them for various purposes, such as prediction of the value of a dependent field (Matsatsinis et al., n.d.).

After the rules are created and their interestingness is measured there is also a call for performing prediction with the rules. Each rule by itself can perform prediction; the consequent is the target and the accuracy of the rule is the accuracy of the prediction. But because rule induction systems produce many rules for a given antecedent or consequent there can be conflicting predictions with different accuracies. This is an opportunity for improving the overall performance of the systems by combining the rules. This can be done in a variety of ways by summing the accuracies as if they were weights or just by taking the prediction of the rule with the maximum accuracy (Thearling et al. ,n.d).

3.5.1.3.2. Rule Induction Algorithms

An induction algorithm takes as input specific instances and produces a model that generalizes beyond these instances. There are two major approaches by induction algorithm: supervised and unsupervised. In the supervised approach, specific examples of a target concept are given, and the goal is to learn how to recognize members of the class using the description attributes. In the contrary, unsupervised approach is provided with a set of examples without any prior classification, and the goal is to discover underlying regularities or patterns by identifying clusters or subsets of similar examples (My Data Mining Weblog (MDMW), 2008).

This research is deal with supervised induction algorithm methods because of simplicity and the goal of the research is prediction about the labeled or known class.

Rule induction seeks to go from the bottom up and collect all possible patterns that are interesting and then later use those patterns for some prediction target. It also retains all possible patterns even if they are redundant or do not aid in predictive accuracy. Hence, in a rule induction system if there were two columns of data that were highly correlated (or in fact just simple transformations of each other) they would result in two rules (Thearling et al., n.d).

Rule induction is also known as **Separate-And-Conquer method**. This method apply an iterative process consisting of first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process (MDMW, 2008). Some examples of these kinds of systems which are supported by WEKA software are discussed below:

1) **OneR:** OneR or "One Rule" is a simple algorithm proposed by Holt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its 'one rule'. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random (Holte,1993).

2) **PART:** PART is a separate-and-conquer rule learner proposed by Eibe and Witten (2005). The algorithm producing sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning (Frank and Witten, 1998)

3) **Decision Table:** It is the method used to build a complete set of test cases without using the internal structure of the program in question. In order to create test cases we use a table to contain the input and output values of a program. It summarizes the dataset with a 'decision table' which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table (Kohavi,1995).

3.5.1.4. Support Vector Machine

A support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis (Vapnik ,1995.) SVM was first proposed and developed by Vapnik (1995).It was gained popularity Piroozniam and Deng (2006) due to many promising features such as:

- Better empirical performance than other classification technique
- The algorithm has a strong theoretical foundation, based on the ideas of VC (Vapnik Chervonenkis) dimension and structural risk minimization
- **Scale up:** the SVM algorithm scales well to relatively large datasets
- **Robustness:** the SVM algorithm is flexible due in part to the robustness of the algorithm itself, and in part to the parameterization of the SVM via a broad class of functions, called kernel functions. The behavior of the SVM can be modified to incorporate prior knowledge of a classification task simply by modifying the underlying kernel function and
- **Accuracy** the most important explanation for the popularity of the SVM algorithm is its accuracy. The underlying theory suggests an explanation for the SVMs excellent learning performance, its widespread application is due in large part to the empirical success the algorithm has achieved.

Support Vector Machine is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. The algorithm performs discriminative classification, learning by example to predict the classifications of previously unclassified data (Piroozniam and Deng, 2006).

The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

3.5.1.4. 1. Method of SVM Classification

The original SVM classification attempts to find the best separating hyper plane with the largest margin width and the smallest number of training errors, as depicted in Figure 3.6

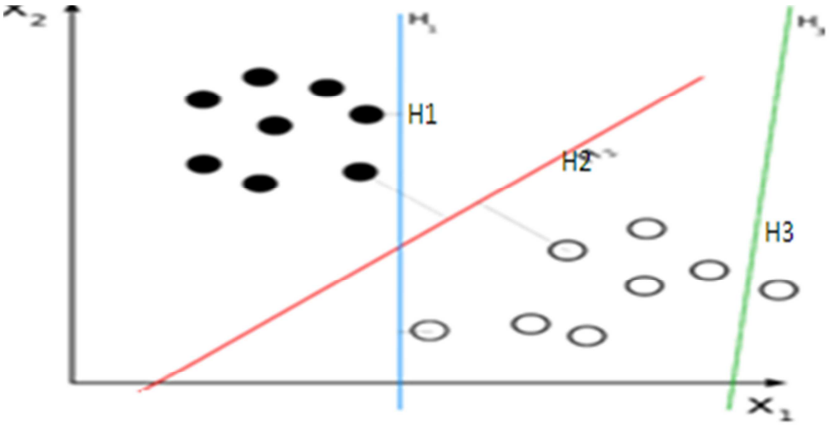


Figure 3. 6 Visualization of linearly separable problem with marginal line

As we can observe in the above figure H3 doesn't separate the two classes. H1 does, with a small margin and H2 with the maximum margin. Besides, it is easy to imagine that the hyper planes that can separate the training data perfectly, the one that has the maximum margin width has a better chance to give the best generalization performance on future unseen data. For this reason, SVM carry out the model of recognition for two-class problems by determining the hyper plane of separation with the maximum distance to the narrowest points of the positioning of formation. These points are called the vectors of support (Chang, et al., 2010).

The learning algorithm of SVM is figure out the boundary that separates the two different classes of training data, subject to certain optimization criteria. Therefore, the exact profile of the boundary is determined by the training instances located in the proximity of the boundary. Consequently, those training instances that are far away from the boundary essentially play no role in determining the boundary. In SVM such a hyper plane with maximum margin is determined by solving a convex quadratic programming (QP) problem. In this section, we present from a mathematical point of view, only the model applied in this study, as the case where the data are linearly separable. According to Han and kamber (2006) discussion the algorithms of Linear SVM work for linearly separable data sets.

Let the data set D be given as (X1, y1), (X2, y2)…, (X |D|, y|D|), where Xi is the set of training tuples with associated class labels , yi.

Each yi can take one of two values, either +1 or-1 (i.e., yi ϵ (-1, +1), for this research case corresponding to the classes HIV = -Ve and HIV =+VE respectively.

There are an infinite number of separating lines that could be drawn to separate all of the tuples of class +1 from all of the tuples of class-1.However, finding the "best" one, that is, one that (we hope) have the minimum classification error on previously unseen tuples is under research.

To solve the stated problem and to illustrate the separating line we use the term "hyper plane" to refer to the decision boundary that we are seeking, regardless of the number of input attributes. The application of SVM approaches this problem by searching for and suggesting for the maximum marginal hyper plane. Consider Figure 3.7 which shows two possible separating hyper planes and their associated margins.

There are an infinite number of (possible) separating hyper planes or "decision boundaries". According to the above figure 3.7 both hyper planes can correctly classify all of the given data tuples. Intuitively, however, we expect the hyper plane with the larger margin to be more accurate at classifying future data tuples than the hyper plane with the smaller margin.

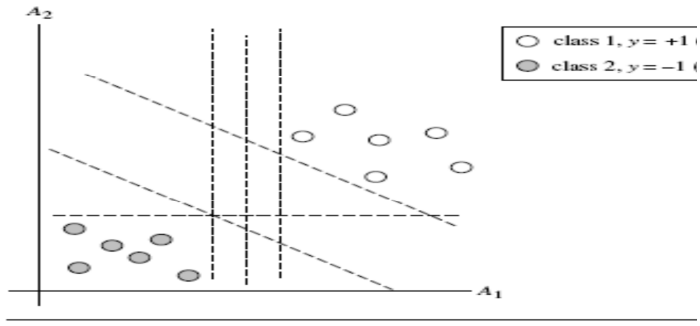


Figure 3. 7 linearly separable 2-D training data

This is why (during the learning or training phase), the SVM searches for the hyper plane with the largest margin, that is, the maximum marginal hyper plane (MMH). The associated margin gives the largest separation between classes. Getting to an informal definition of margin, we can say that the shortest distance from a hyper plane to one side of its margin is equal to the shortest distance from the hyper plane to the other side of its margin, where the "sides" of the margin are parallel to the hyper plane. When dealing with the MMH, this distance is, in fact, the shortest distance from the MMH to the closest training tuple of either class.

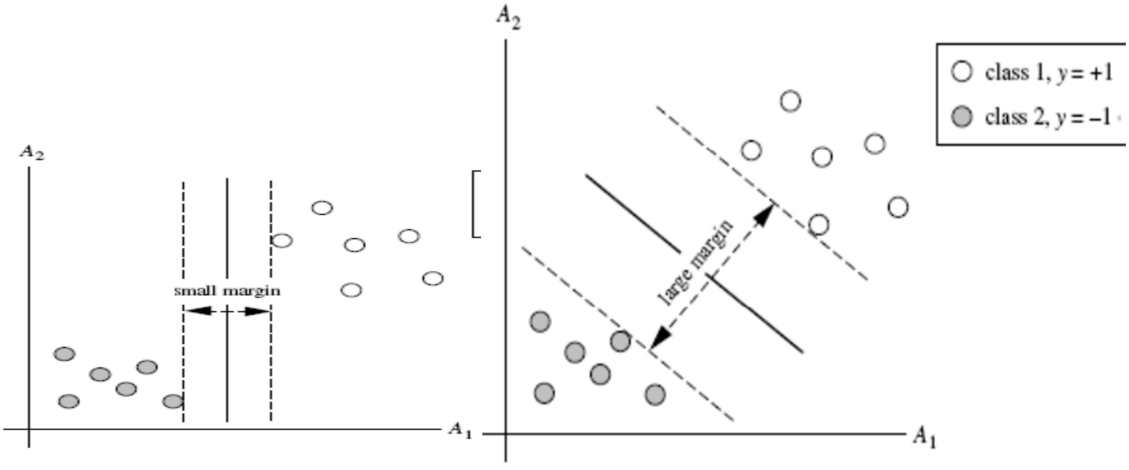


Figure 3.7 (a) (b)

2.5.1.4. 2.Constructing of Optimal Hyper plane Algorithms

Han and Kamber (2006) also reported that how the optimal separating line be drown based on mathematical perspectives. As a result a separating hyper plane can be written as

W*X+b = 0…….. ……..….(3.1)

Where W is a weight vector, namely = {w1, w2… wn}, n is the number of attributes and b is a scalar, often referred to as a bias. Besides, let's consider two input attributes, A1 and A2, as in Figure 3.7(b). The training tuples in 2-D, e.g., X = (x1, x2), where x1 and x2 are the values of attributes A1 and A2, respectively, for X. If we think of b as an additional weight, w0, we can rewrite the above separating hyper plane as

w0+w1x1+w2x2 = 0……..… (3.2)

Figure 3.7 tries to show two possible separating hyper planes and their associated margins. Which one is better? The one with the larger margin should have greater generalization accuracy. Thus, any point that lies above the separating hyper plane satisfies

w0+w1x1+w2x2 > 0…….…. (3.3)

Similarly, any point that lies below the separating hyper plane satisfies

w0+w1x1+w2x2 < 0……...(3.4)

The weights can be adjusted so that the hyper planes defining the "sides" of the margin can be written as H1 : w0+w1x1+w2x2 >= 1 for yi = +1,…….(3.5)

and H2 : w0+w1x1+w2x2 <= - 1 for yi = -1……(3.6)

That is, any tuple that falls on or above H1 belongs to class +1, and any tuple that falls on or below H2 belongs to class -1. Combining the two inequalities of Equations (3.5) and (3.6), we get yi(w0+w1x1+w2x2) >= 1, ∀i……. (3.7)

Any training tuples that fall on hyper planes H1 or H2 (i.e., the "sides" defining the margin) satisfy Equation (3.7) and are called support vectors. That is, they are equally close to the (separating)

MMH. In figure 3.7(b) the support vectors are shown encircled with a thicker border. Essentially, the support vectors are the most difficult tuples to classify and give the most information regarding classification.As Han and kamber (2006) stated linear SVMs we studied would not be able to find a feasible solution here for linearly inseparable data. But there is an approach to extend the steps described for linear SVMs with little modification to create nonlinear SVMs for the classification of linearly inseparable data. Such SVMs are capable of finding nonlinear decision boundaries (i.e., nonlinear hypersurfaces) in input space.

As a result the extended nonlinear SVM methods follows two major steps described below. In the first step, we transform the original input data into a higher dimensional space using a nonlinear mapping. Several common nonlinear mappings can be used in this step. Once the data have been transformed into the new higher space, the second step searches for a linear separating hyper plane in the new space. We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The maximal marginal hyper plane found in the new space corresponds to a nonlinear separating hyper surface in the original space.

3.5.1.4.3. Support Vector machine Algorithms

The Support Vector Machine (SVM) algorithm is probably the most widely used kernel learning algorithm. It achieves relatively robust pattern recognition performance using well established concepts in optimization theory (Cortes and Vapnik, 1995).

There are many kernel functions in SVM, but selecting good kernel function is a research issue. Basically, the followings are some popular kernel functions: linear kernel, RBF kernel, polynomial kernel and hyperbolic tangent kernel. Among this popular kernel functions, the researcher choose the linear kernel function because: the problems are linearly separable, so we don't need other types of kernel function. Besides the class size is two and dataset is easily manageable.

Hence, this research deals with linearly separable data, the researcher used Sequential minimal optimization (SMO) algorithms. Because it is available in WEKA and to implement this we use polynomial kernel function with degree 1 was used .The detail be discussed in experimentation part.

3.5.1.4.4. Sequential minimal optimization

Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines. It was invented by Platt in 1998 at Microsoft Research (Platt, 1998). SMO is widely used for training support vector machines and is implemented by the popular libsvm tool (Chang, C. et al, 2010).

SMO is an iterative algorithm for solving the optimization problem by breaking the problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers α_1 and α_2, the constraints are reduced to:

$$0 \leq \alpha_1, \alpha_2 \leq C \text{…….(3.8)}$$

$$y_1\alpha_1 + y_2\alpha_2 = k \text{ ……...(3.9)}$$

where C is an SVM hyper parameter and K(xi, xj) is the kernel function, both supplied by the user; and the variables are Lagrange multipliers. and this reduced problem can be solved analytically. The algorithm proceeds as follows:

1. Find a Lagrange multiplier that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.

2. Pick a second multiplier and optimize the pair.

3. Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence.

In general, According to Kotsiantis (2007) point out that the following are the general pseudo code for SVM algorithms are depicted as follows.

1. Introduce positive Lagrange multipliers, one for each of the inequality constraints (1). This gives Lagrangian:

$$\boldsymbol{l_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (x_i \cdot w - b) + \sum_{i=1}^{N} \alpha_i} \ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

2. Minimize Lp with respect to w, b. This is a convex quadratic programming problem.

3. In the solution, those points for which α i are called "support vectors" > 0

Even though the maximum margin allows the SVM to select among multiple candidate hyper planes, for many datasets, the SVM may not be able to find any separating hyper plane at all because the data contains misclassified instances. The problem can be addressed by using a soft margin that accepts some misclassifications of the training instances (Veropoulos et al. 1999). This can be done by introducing positive slack variable Si, i =1,..., N in the constraints, which then become:

w.xi – b≥ $+1 - \epsilon \; for \;\; yi = +1$…….(3.11)

w.xi - b≤ $-1 + \varepsilon \;\; for \;\; yi = -1$……(3.12)

$\varepsilon \geq 0$, Thus, for an error to occur the corresponding ε_i must exceed unity, so $\sum_{\mu i} \varepsilon_i$ is an upper bound on the number of training errors. In this case the Lagrangian is:

$$lp \equiv \frac{1}{2}\|w\|^2 + C \sum_i \varepsilon_i - \sum_i \alpha_i \left\{ y_{i\,(x_i .\; w - b) - 1 + \varepsilon_i} \right\} - \sum_i \mu i \varepsilon_i \quad \ldots\ldots\ldots \; (3.13)$$

where the μi are the Lagrange multipliers introduced to enforce positivity of the ε_i .Nevertheless, most real-world problems involve non- separable data for which no hyper plane exists that successfully separates the positive from negative instances in the training set. One solution to the inseparability problem is to map the data onto a higher- dimensional space and define a separating hyper plane there. This higher-dimensional space is called the transformed feature space, as opposed to the input space occupied by the training instances.

Above all this research also investigated using SMO algorithms based on the recommendation of the previous researcher about the domain area and to boost and test the performance of the classification technique of SVM against the status of HIV for a given client.

3.5. 2.Clustering

Clustering is another task of data mining in which groups of instances or objects that are more similar belong to the same clusters whereas dissimilar to different clusters (Han and Kamber, 2001). The only difference between classification and clustering is in the case of clustering unlike classification there are no predefined classes. As there are no predefined classes and examples in clustering, records are grouped together on the basis of similarity among the instances (Beaza-yates and Ribeiro-Neto, 1999).

3.5.3. Association rule

Discovers frequently occurring item sets in a database and presents the patterns as rules. This technique has been applied in problems like network intrusion detection to derive association rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks (Lee et al., 1999).Association learning is a data mining scheme in which any affinity grouping between features is sought, not just one that predict a particular class value (Agarwal and Srikant, 1994).

Association is suitable if the problem is to extract any structure that exists in the data at hand. The aim of association is to examine which instances are most likely to be grouped together. Market basket analysis is a typical practical application of association, which is concerned with what items are consumed with a particular item by individual buyer. Association rules can be developed in order to determine arrangements of items on store shelves in a given supermarket so that items often bought together be found arranged closer in location (Berry and Linoff, 1997).

3.6. Successful Data mining

The two basic things to be successful in data mining is to come up with the a precise formulation of the problem you want to solve and using the right data. The more the model builder can play with the data, build models, evaluate results, and work with the data some more in the given time, the better the resulting model become. Consequently, the degree to which a data

mining tool supports this interactive data exploration is more important than algorithms it uses (Two crows Corporation, 2005).

3.7. Potential Application of data mining

The application of data mining spans various industries. Telecommunications and insurance industries make use of data mining techniques to detect fraudulent activities. In medicine, data mining is used to predict the effectiveness of surgical procedures and medical tests.

Companies in the financial sector use data mining to determine market and industry characteristics as well as to predict individual company and stock performance, Predict which customers buy new policies ;Identify behavior patterns of risky customers (Two Crows Corporation, 1999).

3.8. Review of Related Work

There are a number of researches done to apply data mining techniques in health care domain in general and HIV/AIDS control and prevention program by extracting interesting pattern and rule in supporting the decision maker in particular.

3.8.1. Application of data mining in Health care Industry

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation (Palaniappan and Awang,2008).As a result, Medical institution may use the technologies developed in the new interdisciplinary field of knowledge discovery in databases (KDD) (Larvac, 1998), and particularly data mining (Lloyd-iams, 2002). Today, numerous health care organizations are using data mining tools and techniques in order to raise the quality and efficiency of health-related products and services. A number of articles published in the health care literature shown the practical application of data mining techniques in the analysis of health care related records.

As part of this effort the health care organizations are implementing data mining technologies to help minimize costs of clinical test, provision of quality service (diagnosing patients correctly and administering treatments that are effective) at affordable cost and improve the efficiency of patient

care. Besides, data mining can be used to help predict future patient behavior and to improve treatment programs.

By identifying high-risk patients, health professional can better manage the care of patients so they will not be problems of tomorrow (Rogers and Joyner, 2002).Recent trends in medical decision making show awareness of the need to introduce formal reasoning, as well as intelligent data analysis techniques in the extraction of knowledge, regularities, trends and representative cases from patient data stored in medical records (Larvac,1998).

According to Larvac, machine-learning methods have been applied to a variety of medical domains in order to improve medical decision-making. Razak and baker (2006) have also investigated a study focusing on mining association rules from asthma patients profile dataset. The purpose of the study was to identify attributes that affect asthma patients. The dataset of the asthma patients records were about 16,384 records and 118 variables in various formats. These attributes are grouped into demographic attributes and asthma related attributes.

They employ data preparation and association rule mining phases of data mining method. Understanding the nature of the dataset, identifying data types and formats, identify incomplete data, analyzing data distribution, and discretizing data are the important stages needed in order to systematically preprocess the data. As result of data preprocessing and cleaning only 31 attributes are left to generate association rules. The association rules mining phase uses Apriori Algorithm. To implement the association rule mining technique, testing and training dataset by determining threshold value is the main work of the researcher.

Palaniappan and Awang (2008) have investigated that how to predict the probability of patients getting a heart disease given patient records. In order to achieve this data mining goals, they used CRISP-DM methodology and data mining Extension (DMX) query language for graphical user interface design. They developed an intelligent system that predicts a given patient heart disease status. To develop a model the study used decision tree, naïve- byes and Neural network data mining techniques. The study used a total of 909 records with 15 medical attributes are used. The records are split equally into training dataset (455 records) and testing dataset (454 records). For

the sake of consistency, only categorical medical attributes are transformed to categorical data. The attribute diagnosis" is identified as the predictable attribute.

The prototype intelligent heart disease prediction system (IHDPS) developed based on the combined output of decision trees, naïve Bayes and neural network data mining techniques. The results of the study are encouraging and realize the objectives of the defined mining goals. The system extracts hidden knowledge from a historical heart disease database. The models are trained and validated against a test dataset. All model to have extracted patterns in response to the predictable state. The registered performance of naive byes is better to predict patients with heart disease and followed by neural network and then decision trees.

3.8.2. Application of data mining in HIV/AIDS

Various local and international scholars have attempted to study the application of data mining on HIV related records or data. The following works are reviewed by the researcher to help clarify the significance of the research problem and to show the gap in the current local and international studies.

Vararuk et al. (2008), for example, have investigated the application of data mining techniques on HIV/AIDS records to determine the pattern that exists in patient. The researchers patterns obtained in their study can be used for better management of the disease and appropriate use of resources in Thailand. The study took 250,000 records from HIV/AIDS patients. IBM's intelligent miner is used for clustering and association rule mining is applied to identify symptoms that may follow a set of existing ones.The study result of Vararuk et al.(2008) have showed clustering of group of patients with common characteristics . Association rules generate that were not expected in the data and are different from traditional reporting mechanisms utilized by medical practitioners. It also allowed the identification of symptoms that co-exist together. A good indicator of this research was that identification of symptoms that are forerunner of other symptoms can allow the targeting of the former so that the latter symptom can be avoided.

Similarly, Madigan, et al. (2008) have examined the association between the health workforce, particularly the nursing workforce, and the achievement of the millennium development goal, taking into account other factors known to influence health status, such as socioeconomic

indicators. Specifically, the main goal of the study was to investigate the association between health workforces with HIV prevalence rate at country level (worldwide). The dataset obtained from WHO and NGO sources are merged into one file at the country level for analysis. CART Methods over clustering data mining technique is used for prediction of the levels of HIV/AIDS prevalence rates among the 194 countries for which data are collected. The predictive model developed indicates that increased physician and nurse density (number of physicians or nurses per population) was associated with lower adult HIV/AIDS prevalence rate, even when controlling for socioeconomic indicators. Finally the performance of the model revealed that the main determinate factors in understanding HIV//AIDS prevalence rates are physician density followed by female literacy rates and nursing density in the country.

Elias (2011) conducted his research to determine the status of clients being HIV positive and negative for data collected from Addis Ababa Administration HCT users. To predict client's status the researcher used decision tree, J48 algorithms and Association rule of Apriori algorithms. To develop a model he followed CRISP –MD methodology. For preparation of dataset Weka 3.7.5 , and Microsoft Excel was used. The research output indicated that 81.8% performance was registered to associate the attribute and HIV status. Finally, the researcher recommends the implementation of the same research problem with other data mining techniques by increasing the number of dataset, number of variable to register better performance and develop knowledge based system based on the extracted rule to scale up HIV testing in Addis Ababa.

Teklu (2010) has attempted to investigate the application of data mining techniques on anti-retroviral treatment (ART) service with the purpose of identifying the determinant factors affecting the termination /continuation of the service. This study applied classification and association rules using, j48 and apriori algorithms respectively. The study was conducted using 18,740 ART patients' dataset. The methodology employed to perform the research work is CRISP_DM. Finally, the investigator proved that the applicability of data mining on ART by identifying those factors causing the continuation of the service. Birru (2009), has also investigated the applicability of data mining on VCT taking the case of CDC. He used 56,486 dataset from 2002 to 2008. The dataset contains unbalance HIV positive and negative clients' data and after the dataset is balanced only

14,793 records are considered for his experiment. To develop a model he employed CRISP-DM methodology and used clustering and classification data mining techniques.

Among Clustering techniques, k-means and Expectation maximization (EM) algorithms are used to define group of similar VCT client and to see how these grouped affect the classification outcome. From the two clustering algorithms EM is selected and the cluster indices created using this algorithm is then used as class for the classification purpose. In implementing the classification he has used J48, random tree and multi-layer perception to predict level of risks of clients as high risky or low risk based on the clustered indices. The performance of the model indicate that decision tree have shown better performance and appropriate to the domain. This is due to the fact that decision tree algorithm has a simple feature which can be easily understanding by non-technical staff. Above all the researcher recommended further research using large dataset and other data mining technique to boost the performance of the model.

Correspondingly, Abraham (2005) has conducted data mining research that examine the application of data mining technology to identify determinant factor of HIV infection and to find their association taking the case of CDC. The researcher used 18,646 records extracted from database of CDC. This dataset consists of 82 attribute, among them only 19 attributes were selected for training and testing the model. To develop a model the researcher used decision tree and association rule discovery. Knowledge STUDIO and Weka 3.7.5 data mining tools are used to implement the model and to extract rules to identify risk factor. Abraham has tried different association rule mining experiments by reducing attributes which the algorithm can do. He has attempted to find risk factors using only general association rule which it can bring any attribute in the consequent of the rule which may not be useful to identify most risk factors.

In general, the aforementioned review highlight that much has not been done and left out recommendation to try by other researcher in way to improve the performance of the model and get the actual problem solved. In this regards, this research has been attempted to determine variables that has significance impact on HIV infection. Consequently, the researchers have used 20 attribute, three different classification model, and different target population (no previous

studies has investigated the behavior of University community towards HIV with this technology) to attain the research objectives.

Hence, to the knowledge of the researcher, this research problem has not been attempted before and novel because more variables are used for training and testing the model; some of the features of the variables or attributes are different (such as location, residence and category) and lastly some of the classification model such as Support vector machines are not been used yet to investigate the determinate risk factor of HIV infection.

Taking this fact in mind, this research has initiated and conducted ,and would be expected to contribute a lot to managers and decision makers in HIV/AISD prevention and control practicing in VCT center.

CHAPTER FOUR

DATA UNDERSTANDGING AND PREPROCESSING

4.1. Overview of African AIDS Initiatives International

AAII (African AIDS Initiative International) is a Non profitable and Non-Governmental Organization (NGO) and is registered both in the State of Massachusetts, U.S.A and Ethiopia working with the motto of "Fighting for Life". (LLGP, 2007).

AAII envisions HIV/AIDS free, healthy and productive Africans. The mission of the organization is preventing the incidence of HIV infection and mitigating the impacts of AIDS among Africans. The goal of the organization is developing and promoting a culture of excellence for HIV/AIDS prevention, service delivery, networking and research.

To achieve its goal, the organization aims at providing HIV/AIDS education for positive behavior change, creating access to Voluntary HIV/AIDS Counseling and Testing (VCT) services, provide care, support and treatment for people living with HIV/AIDS (PLWHA) and Orphans, spearheading and facilitating collaborative partnership among the national and international actors working on HIV/AIDS in Africa. AAII opened its Regional Office for Africa in Addis Ababa in April 2001. It is strongly believed that the AAII's Ethiopia operations will serve as a model to replicate its activities in other African countries in collaboration with African Union (AU), Harvard AIDS Institute and World Health Organization (WHO) (LLGP, 2007).

As part of its effort to contain the spread and mitigate the impact of HIV/AIDS at national level, AAII is involved in various HIV/AIDS prevention and control activities. In this regard AAII is involved in the provision of VCT services, support to students' and workers' associations, girls' and reproductive health (RH) clubs, provision of quality HIV/AIDS/RH services through establishing resource centers, care and support to PLWHAs in the various faculties and schools of Addis Ababa University. Apart from its activities in the academic institutions, AAII is involved in providing care and support to PLWHAs and AIDS orphans in Mojo, Debre Birhan, Debre Markos

and Addis Ababa. Besides, it provides support to commercial sex workers in selected areas of Addis Ababa, particularly (Arada Sub City).The intention of this research is to evaluate the practice of the university community towards HIV infection insides and outsides university(area where university community exposed to their personal entertainment) campus .

As a result, this research mainly focuses on VCT service provided in Addis Ababa university community and centers that resides in Arada Sub city; particularly centers that are found in the surrounding of Addis Ababa University. The VCT service provided in different faculty mainly in Arat kilo and Sedist Kilo campus and off campus center which is located in Arada sub city around Minillek Hospital. The process of VCT service is discussed in chapter three, section 3.4.1 in detail. The counsellor provides both pre-testing and post testing counselling. According to the counsellor of AAII they spent much time in pretesting phase in discussing issue related to HIV/AIDS. Once the client decided to get test the counselling process takes place in the following ways.

They initially fill client information in standard paper format manually see **Appendix 1.** The format has three parts, the first part will be filled by receptionist, the second part called "pre-testing" which is filled by the counsellor and the last part called "post testing" filled by counsellor too. After pre-testing phase, the counselor takes the client to the laboratory for test if and only if the client decided to take the HIV test. The information gathered from clients is collected from all VCT centers to be encoded centrally in computerized system. The software is called Epi Info system which is capable of handling medical related VCT data. The documentation, data entry and analysis are performed in the data management section. This section is located in off campus site and has one IT (Information Technology) professional to encode data and handle manipulation of the system.

The proximity and being staff of the University, the researcher's get access to and collect the required data for this study encouraged the researcher to conduct this research at AAII.

4.2. Business Understanding

The main goal of AAII is the realization of an HIV/AIDS free community at all operational level sites where men and women in general and youth in particular are empowered to use their potential to control the spread and mitigate the impact of HIV/AIDS. AAII achieves its objective

empowering youth through enhancing the knowledge, attitude/behavior and life skills of the student to detect (high) risk activates that leads to acquisition of and transmission of the HIV pandemic. Besides, expansion of VCT service in campus as well as off campus site is the main organizational goal. According to the counsellor and data manager of the AAII, in campus site various means are used to teach the university community to mitigate the spread of HIV/AIDS in campus.

These are conducting consecutive training for trainer to the selected student member, distribution of condom, and leaflet through student representative across dormitory, student lounge and within their office. Besides, pear education is the best mechanism that the organization is currently using to create awareness about the incurable disease. This mechanism basically works with one student for a minimum of five students (1:5) is assigned and the assigned student is responsible for introducing about the disease for his/her specific group. She/he can be consider as a "preacher" and as a result s/he gets a pocket money for accomplishing the assigned task.

On the other hand, the organization also provides voluntary counselling and testing service in off campus site to address the service to university surrounding and commercial sex worker community.

4.2.1. Classification of AAII intervention area

AAII just like other international as well as local organization has its own five years strategic plan to accomplish its mission. It is a licensed and registered international organization. Its license is renewed yearly based by Ministry of Health. It has established a framework to be followed under five major intervention areas. These are:

1. **Prevention education:** Providing HIV/AIDS education to create awareness and bring behavioral change. Providing financial assistance for student anti-AIDS club and worker and female students club.

2. **Condom promotion and distribution**

3. **Voluntary counselling and testing**: to create the opportunity and access to quality VCT services. In this regards AAII establish four VCT sites: Sidist kilo, Science, Mobile VCT

center and Off campus(to provide access to those who are not comfortable to use in campus service)

4. **Care and support**: It also provides financial assistance especially for sex worker and economically unaffordable client. Because of lack of sufficient fund the AAII set up an agreement with university surrounding government hospital such as Black lion, yekatit 12 and Menilik for providing ART(Antiroviral Treatment) service.

5. **Capacity building and Networking and research.** Establish database and network facilities for disseminating information. Facilitating a networking and partnership among the VCT community.

To accomplish its task the main source of fund AAII is from USAIDS, UNAIDS, Care International, HAPCO (HIV/AIDS Prevention and Control Office) and Federal HAPCO**.** Even though, such fund are earned through competition of project proposal the AAII has the following three major problems.

a) Renewal of licenses in annual base is too bureaucratic and time taking. As a result of this the organization missed project funder.

b) Lack of transportation facility such as car.

c) Lack of sufficient budget to expand the scope of VCT service. As discussed above, the organization provides only VCT service, after the client result is positive, it doesn't provides financial support for that client.

In a normal circumstance AAII provides voluntary counselling and testing service to the community in campus and off campus site. The organization only told the laboratory result of clients after being tested in either site. Beyond that there is no way of identifying variable that cause HIV infection rather keeps the client information in computerized system. Basically AAII analyze reports or the activities of the VCT services using traditional methods of data analysis such as MS Excel (mean, mode, variance etc.). However, such methods of data analysis has limited capacity to discover new and unanticipated patterns and relationship that are hidden in

conventional databases (Plate et al.,1997). Besides, those identifying patterns of attribute or risk factors of HIV infection are difficult because the dataset contains or involves too many attribute or parameter. So studying the patterns of the attribute that have relationship among each other rather than listing are more crucial and support decision making.

Consequently, applying DM technology can handle and generate such pattern of variables that cause HIV infection and used the output by concerned party to facilitate decision making process. Hence, this is main factor that the researcher initiated to conduct this research.

4.3. Understanding of Data

According to Cios et al. (2007) model, the next phase after understanding the business and problem domain has been data understanding. Hence, a prerequisite for undertaking DM research was dealing with data itself. For this reason, Berry and Linoff (2004) described that the identified good source of data to attain the DM goals has been the corporate data warehouse. Hence the data stored in corporate database (AAII VCT database in this context) maintain consistency, in their attribute value, data types and common format.

4.3.1. Data Source and Data Collection

There are two key tasks in data mining. The first one is coming up with precise formulations of the problem you are trying to solve. The second task is using the right data (Two Crows Corporation, 2006). Of the two main tasks analyzing and understanding the content and structure of the collected data is one of the most important tasks that need attention in data mining process.

The source of data for this research has been collected from AAII VCT center database. The data base was manipulated by using Epi info software system. It was standard medical software that suite for handling medial related data. It has a user interface that allows the data encoder to enter clients' information. There were 18,800 clients records in the database but the researcher used only 15,396 for the reason that some records are noise ,incomplete, and out of the scope of the study. All the records are taken from the database starting from 2005 to 2011 G.C. The summary of the data sources is illustrated below in Table 4.1

Table 4. 1 Description of data source and Number of records

Source of data center	Data coverage	Number of records	Number of attribute	Size of the data	Data Type
AAII VCT data	2005-2011 G.C	15,398	63	4.2MB	Nominal

4.3.2. Description and Quality of Data

The number of attributes of collected data from the VCT center is about 63. However, with the discussion of the domain expert, only relevant attributes is selected to perform data mining tasks. The selected attributes are described in Appendix 2. According to Han and Kamber (2006) data quality can be measured in terms of accuracy, completeness, consistency, timeliness, believability and interpretability. In this regard, VCT data suffer from incompleteness, missing value, invalid value and inconsistency problems. Since data quality can also be affected by the structure of the data being analyzed, the researcher give due emphasis to this aspect of the data too. In the next section only the selected VCT data for experimentation is summarized statistically relative to the whole data.

4.3.3. Descriptive Statistical Summary of Selected attributes

This initial dataset has been described and visualized using Microsoft Excels to examine the properties of the dataset relative to the whole records. Simple statistical analysis has been performed to verify the quality of the dataset such as missing values, error values and to obtain high level information regarding the data mining questions. Hence, the selected attributes used for model building are statistically described in details below. This is helpful for understanding of the dataset for experimentation.

4.3.3.1. Attribute Selection

In this study VCT data has been used, however, the original records were full of problems and errors; they were not used directly for data mining purpose. The process of filling client's information manually consists of three phases. In the **first phase**, there are 19 attributes to be filled

at reception when clients came to the sites for VCT services. These are "**Country", "Region","Woreda", "Site code", "Site type", "Organization type", "Resident", category", "year ", "Campus", "Client code", "Return visit", "New code", "Age", "Sex", "Counselor code", "Partner code", "Couple code", and "Visit date".**

Out of this only **"Year"," Campus", "Residence", "Age" and "Sex"** attributes were selected, because all of them directly related to the problem to be solved to attain the research objectives. However, Country and Regions have been removed because contains single value attribute and has no role for prediction to the subject under study.

Moreover, "Site code", "Site type", and "Organization type "attributes have not been used to identify the type of the centers but have not relevant for this particular problem. Similarly "Client code"," Counselor code and "Partner code" were not used since they are unique values and not relevant in showing patterns. "Return visit" and "New code" attributes were used to check whether the client has been returned to the center or new, but these attribute was redundant instead" previous test attribute" was used.

Furthermore, pre counselling phase consists of 25 attributes (see Appendix 1), out of which 15 of them has been selected. Other attributes such as how you hear of about VCT, from where the client is referred, and date previously tested were not considered as relevant for the purpose of this research. All attributes in the post-test counseling session that are obtained after knowing the HIV status were not considered since they did not help in predicting the HIV infection risk factor of the clients. Thus, 20 attributes are selected for experiments.

The selected final attributes were statistically described below and see the details in Appendix 2.

Residence attributes. It stands for the place where client came from. The attribute has 12 values and they were nominal.

The frequency of the attribute with its various values looks like as depicted in Table 4. 2. A client who came from Addis takes 74.41 % of the whole dataset. The missing taken about 1.64 %, however; this value was filled by the most frequented value which was Addis.

Table 4. 2 Summary of Residence's Attribute

Residence : Nominal		
Distinct values	**Frequency**	**Percentage (%)**
10-Addis	11458	74.41
11-Dire	59	0.38
1-Region 1	720	4.68
2-Region 2	19	0.12
3-Region 3	1178	7.65
4-Region 4	1167	7.58
5-Region 5	14	0.09
6-Region 6	31	0.20
7-SNNP	422	2.74
8-Region 8	16	0.10
9-Region 9	31	0.20
Others	31	0.20
Missing Value	252	1.64
Total	15396	100

Category attribute: The frequency of the attributes and the corresponding possible nominal value has been shown in Table 4.3. As shown in table, the modal value for category was regular student. It has eleven values. The missing value of the attribute was 9867 (64 %). It takes more than half of the total records percentage. According to Han and kamber (2006) applying the most frequent value (here regular student) is used to fill the missing value.

However, when we observe the actual records their residence value is Addis (urban). So discussing with the domain expert the missing value is filled with community. This is because the number of the off campus client is greater than in campus client.

Table 4. 3 Statistical summary of category attribute

Category : Nominal		
Distinct value	**No of clients**	**Percentage (%)**
Regular student	5854	38.02
Extension student	359	2.33
Summer	403	2.62
Staff	967	6.28
Community	845	5.49
Student partner	306	1.99
other	1133	7.36
Missing value	9867	64.08
Total	15396	100.00

Year Attribute: It is a nominal attribute having ten distinct values including 1st year to 8th year and the other which are not applicable is another value. By examining the current university curriculum, the researcher reframes the distinct value as follows in Table 4.4.

As shown in the following Table 4.4 , the common university community who taken VCT service was third year but the community by far greater because their status belong to non - student. It is about 57% of the total population. Basically the initial data indicate that both staff and community were contained a missing and very few with not applicable value.

Based on the reality and convenient for experimentation result the researcher filled the missing value not with modal value rather with not applicable.

Table 4. 4 Statistical description of the year attributes

Year :Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Not applicable	402	2.61
1st year	1326	8.61
2nd year	1399	9.09
3rd year	1725	11.20
4th year	1595	10.36
5th year	123	0.80
6th year	33	0.21
Missing	8795	57.12
Total	15396	100.00

Missing value of the records in the above table registered 57.12 % of the total selected dataset. But, the actual records of the missing value was belongs to non-student and found unfilled (missing) similarly to Table 4.3, missing value is filled with not applicable value instead of 3rd year student which is relatively the most frequent value.

Sex Attribute: It contains two valid values that are either male or female. The detail is presented in table 4.5. As shown in table 4.5, male clients (63.39%) are more than female clients (36.52%).This implies more number of male than female clients was visiting the centers.

Session type Attribute: This attribute indicates whether the clients come to visit the center individually or in couple. It has four possible value namely individual, group, couple and other. Table 4.6 presents the distribution of clients by session type.

Table 4. 5 Statistical summary of Sex attribute

Sex :Nominal		
Distinct value	Frequency	Percentage (%)
Male	9761	63.39
Female	5623	36.52
Missing Value	14	0.09
Total	15396	100.00

Table 4. 6 Statistical summary of session type Attribute

Session Type: Nominal		
Distinct Value	**Frequency**	**Percentage (%)**
Individual	13410	87.09
Couple	1891	12.28
Group	9	0.06
Other	9	0.06
Missing value	79	0.51
Total	15396	100.00

Marital status Attribute: The Attribute of marital status contains seven distinct values such as married, never married, separated, divorced, widowed, living together and other. The frequency of the attribute with possible nominal value will be described in Table 4.7

As shown in the table below, modal value for marital status was never married. The missing values of the attribute were 90 (.58%) which is very little so it better to fill with the modal value.

Table 4. 7 Statistical Summary of Marital status Attribute

Marital status: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Married	1514	9.83
Never married	13187	85.64
Separated	129	0.84
Divorced	290	1.88
Widowed	121	0.79
Live together	7	0.05
Other	60	0.39
Missing value	90	0.58
Total	15396	100.00

Education Attribute: This attribute reveals the level of education of clients. It has nine distinct values (illiterate, PHD, Able to read, Primary, Secondary and other mentioned in following Table 4.8).

Employee Attribute: This attribute has a nominal value and refers to the client's status who employed or not. The distribution of the client about their employment situation has been described in Table 4.8. It has two distinct value such as Yes and No.

Table 4. 8 Statistical Summary of Employee Attribute

Employee: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
0-No	9092	59.05
1-YES	6143	39.89
Missing value	163	1.06
Total	15396	100.00

Table 4. 9 Statistical Summary of Education Attribute

Education: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Illiterate	249	1.62
PHD	3	1.94
Able to read	245	1.59
Primary	1686	10.95
Secondary	3969	25.7
Diploma	1640	10.65
B.Sc.	452	2.94
Undergraduate St.	6366	41.34
Postgraduate St.	382	2.48
BA degree	158	1.03
Other	94	0.61
M.A degree	2	0.01
Missing value	152	0.99
Total	15396	100.00

As shown in table 4.9, clients whom educational status was undergraduate student and visiting the centers most frequently accounts 41.34% of total dataset. It followed by clients who has secondary education level (25.78%) and primary educational levels (10.95%) respectively .Some missing (0.99%) values were observed in AAII VCT data

In the above table we can visualize that the percentage of client being employed was less than form not employed. The percentage of not employed versus employed is 59.05 and 39.89 % respectively. The remaining percentage is missing value which is 163(1.06%).

Primary Reasons Attribute: This attribute describes the reason why the client is visiting the center. There are around 21 distinct values for primary reason for using the VCT service. The frequency of the distinct value with their respective percentage value will be shown in the Table 4.10 below.

Table 4. 10 Statistical Summary of Primary Reasons Attribute

Primary Reasons: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
2nd Test(win)	833	5.41
Confirm positive result	70	0.45
Get results previous test	2	0.01
Need Counseling	453	2.94
Test before pregnant	9	0.06
Pregnant must know	28	0.18
Plan for future	9322	60.54
Death/illness of partner	48	0.31
Occupational exposure	54	0.35
Other blood/Fluid exp.	294	1.91
Client Risky/Had risk	1462	9.49
Sexual assault	14	0.09
To know status	70	0.45
Partner Risky/Had risk	180	1.17
Not trust partner	397	2.58
Ill/Symptoms	174	1.13
Premarital	409	2.66
Marital reunion	31	0.20
Family planning	3	0.02
Visa application	142	0.92
Other	1120	7.27
Referred	7	0.05
Missing value	**245**	**1.59**
Total	**15396**	**100.00**

As shown in table 4.10, most of the clients who visiting the centers for planning their future (which accounts for 60.54%) followed by when their client was at risky condition (9.49%). Moreover, 245(**1.59**%) missing values were identified from the data.

Previous Tested Attribute: This attribute refers to whether clients have been previously tested or not. The clients visiting the centers may have come previously and have different previous HIV status or they may be coming for the first time. The summary of the previous test status of the clients is presented in table 4.11**.**

Table 4. 11 Statistical Summary of Previous tested Attribute

Previous Test result: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
No	5968	38.76
Yes, HIV+	204	1.32
Yes, HIV -	8998	58.44
Yes, inconclusive	13	0.08
Result not given	2	0.01
Didn't take results	5	0.03
Other	13	0.08
Missing values	195	1.27
Total	15396	100.00

Table 4.11 indicate that even though, there have been various reason for getting HIV test, getting negative result encouraged clients to come again in the VCT Center (Which accounts 58.44%) and followed by clients not previously tested (it accounts 38.76 %).

Ever Had Sex Attribute: This attribute shows whether the client has ever had sex since s/he was born or not**.** The summary of the data are presented in table 4.12

Table 4. 12 Statistical Summary of Ever Had Sex Attribute

Ever Had Sex Type: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
No	5525	35.88
YES	9734	63.22
Missing Values	139	0.90
Total	15396	100.00

As shown in table 4.12, most clients (63.22%) had experienced sex when they come to visit the centers. There is 0.90% missing values.

Suspected Exposure Time: This attribute is the length of time between the date of the client's most recent possible exposure to HIV and the date of the VCT visit. Table 4.13 shows the detailed distribution of clients by exposure time.

Table 4. 13 Statistical Summary of Suspected Exposure Attribute

Suspected exposer Time: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
No exposure	3708	24.08
<1 month	1298	8.43
1 to 3 months	1209	7.85
4 to 6 months	2035	13.22
Over 6 months	2998	19.47
Don't know	2883	18.72
N/A	1093	7.10
Missing values	174	1.13
Total	**15396**	**100.00**

As shown in table 4.13, the suspected exposure time accounts of 24.08% clients who has no exposure and followed by Over 6 months (which accounts 19.47%) when they visit the center. Moreover, 1.13% missing values is detected from the data.

Condom Use Attribute: It refers to the condom usage manners of clients during sexual intercourse for the last 3 months. The frequency distribution of clients are illustrated in table 4.14

Table 4. 14 Statistical Summary of Condom Use Attribute

Condom use: Nominal		
Distinct value	**Frequency**	**Percentage**
Never	5236	34.00
Always	2134	13.86
Sometimes	770	5.00
N/A	7081	45.99
Missing values	177	1.15
Total	**15396**	**100.00**

As illustrated in the above table 45.99% of the clients are observed as not applicable for use of condom and those who accounts about 34% are never use the condom. Additionally, 177(1.15%) of the client use condom behavior for the last six month is treated as missing value or unfilled.

Used condom Last Sex Time attribute: This nominal attribute is talk about the client's condom use behavior during last sexual intercourse with sexual partner. It has four possible distinct value namely: - yes, no, does not remember and N/A (not applicable). The frequency distribution is depicted be in Table 4.15

According to Table 4.15 indicated that 46.7 and 31.79 % of the clients have not used condom during last time they had sex, however 19.74 % have used condom. Minimum number of missing value (1.09%) is identified from the records.

Table 4. 15 Statistical Summary of Used Condom last Attribute

Condom use Last Time had sex: Nominal		
Distinct value	**Frequency**	**Percentage**
No	7191	46.70
Yes	3040	19.74
Doesn't remember	104	0.68
N/A	4895	31.79
Missing value	168	1.09
Total	15396	100.00

History of Sexual Transmitted Infection (HSTIs**) attributes:** This attribute deals about client's personal history of has been make a diagnosis or practice of STIs. It has same distinct value of as "condom use last time had sex" attribute. The detail is presented in table 4.16.

Table 4. 16 Statistical Summary of History of Sexual Transmitted infection Attributes

H STI: Nominal		
Distinct value	**Frequency**	**Percentage**
No	10263	66.65
Yes	150	0.97
Don't know	13	0.08
N/A	4769	30.97
Missing value	203	1.32
Total	15396	100.00

In the above table we can observe that 66.65% of the clients have not known their sexual transmitted infection history. The number of missing value is very few and accounts about 1.32%.

Steady Attribute: A numeric attribute that indicate the number of permanent sexual partner that clients have ever contacted with. It described in terms of mode, mean, outlier and missing value. Statistical distribution of the number of sexual partner for a given client be illustrated below in Table 4.17.

Table 4. 17 statistical summary of steady attribute

Distinct Value	Frequency	Percentage(%)	Remarks
Invalid value	**6261**	**40.66**	**1-,-01,-,_,**and *
0	1990	12.92	
1	6969	45.26	
2	18	0.12	
3	4	0.03	
4	8	0.05	
5	1	0.01	
6	1	0.01	
7	**1**	**0.01**	
10	**3**	**0.02**	
14	**5**	**0.03**	
Missing value	**137**	**0.89**	
Total	**15396**	**100.0**	

Table 4.17 shows that the number of client who has permanent partner account of about 45.26% and followed by those who have no partner or zero value. The number of error or outlier and missing value detected from the data is 40.66 and 89% respectively. On the other hand almost half of the records were filled by invalid value.

Age attribute: The age attribute contains a numerical valid value in the range from birth to 99 years old client. Clients at various ages are coming to the centers for HIV screening. The details of the age of the clients are presented in table 4.18. As the researcher observed, there is no age missing value from the selected dataset.

More over the average age who involved in VCT service is 25. This indicates that they were the youngest and productive force of the community. Hence, the dataset includes both age groups

starting form 0 year to older age because the organization provides service for both university community and people who are found in surrounding of Addis Ababa University. The maximum age of client who took the VCT service is 85.

Table 4. 18 Statistical summary of age attribute

Age : numeric		
S,N	**Age distribution**	**Age number**
1	Minimum	0
2	Maximum	85
3	Mode	20
4	Average	25
Total number of Age in the dataset		**15396**

Location /Campus attribute: This nominal attribute records the client place of work or their affiliate faculty in which they registered. It is classified under nomenclature of the university faculty and arada sub city. The frequency of the clients, where they are living ,and where they takes VCT service depicted below in Table 4.19.

As shown below in table 4.19 majority of the client who took VCT service have not mentioned their campus or place where they currently settled. Which account about 55.26% considered as a missing value. But about 28.07 % of the client was living in sedist kilo. The some value of the attribute duplicated with different spelling of same name. For instance, ‘Art’ and ‘art school’, registered with different name but same value. It was corrected and categorized as one distinct value.

Table 4. 19 Statistical Summary of place of location/ campus of the client

Campus/Location : Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Art	1	0.01
Pharmacy	1	0.01
Sedist kilo	4322	28.07
Science	2134	13.86
N.Technology	129	0.84
S.Technoligy	32	0.21
Commerce	21	0.14
Veterinary	5	0.03
B.Lion	4	0.03
ART School	3	0.02
Yarede	2	0.01
N/A	51	0.33
Other	83	0.54
B.Lion	1	0.01
Missing Value	8509	55.26
Total	**15396**	**100.00**

Similarly, as Table 4.3 and 4.4, here is a missing value which was handled using other than most frequented value. Because the campus only hold the value of the student rather than community or client. When we observed the actual data set the missing value belongs to "Addis". So the researcher with much discussion with domain expertise decided to fill using NA value.

Causal Attribute: This numeric attribute shows the number of all those sex partners the client has faced only once or rarely for the last 3 months. Table 4.20 summarizes the frequency distribution of the attribute causal. As shown below in table 4.20 the number of erroneous treated as causal value is 69.21%. The missing value is 1.07%. The average number that a client faced a sexual partner is less than one which is account of .95 % and it is followed by usual one partner contact.

Table 4. 20 Statistical summary of Causal Attributes

Causal: Numeric			
Number of Causal	**frequency**	**Percentage (%)**	**Remark**
0	1979	12.85	Minimum contact------0 Maximum contact------20 Average----------0.95 Missing value----------1.07% Error----------69.21%
1	2005	13.02	
2	302	1.96	
3	194	1.26	
4	43	0.28	
5	20	0.13	
6	11	0.07	
7	6	0.04	
8	1	0.01	
10	6	0.04	
11	2	0.01	
12	2	0.01	
15	3	0.02	
18	1	0.01	
20	1	0.01	
Missing value	165	1.07	
Invalid value	10657	69.21	--,-,-*,*,-+,0-,,1-,4--,4- are outlier
Total	**15396**	**100.00**	

This takes 13.02 % of the total client. Besides, the outlier that is wrongly entered in the record should be removed and corrected manually based on the proximity of the value by the researcher judgment.

HIV Test Result Attribute: This nominal attribute determine the client HIV status by laboratory test of his/her blood. The laboratory test will be investigated and determine by lab technician as HIV negative and positive. The overall investigation of the AAII data set regarding HIV status is illustrated in Table 4.21.

As depicted in the table 4.21 about 96.10 % of the client has be tested and grouped as HIV negative, 3.07 % of the clients are HIV positive. The number of missing value and error detected from the data are .79% and 0 .01% respectively.

Table 4. 21 Statistical summary of HIV Test Result Attributes

HIV test: Nominal		
Distinct value	**Frequency**	**Percentage**
Negative	14798	96.10
HIV+	473	3.07
NA	4	0.03
Invalid	1	0.01
Missing value	122	0.79
Total	**15396**	**100.00**

4.4. Data Preparation and Preprocessing

These steps consist of the core of data mining and take much of the research time and effort of the entire knowledge discovery process but critically, the most compulsory activity in data mining process (Han and Kamber, 2001).

Data preprocessing involves all the action taken before the actual data analysis process start. It can be defined as a transformation that transforms the raw real world data to a set of new data (Famili and Turney, 1997). It is unrealistic to expect that data will be perfect after it is extracted. Since good models usually need good data, a thorough cleansing of the data is an important step to improve the quality of data mining methods.

As Berry & Linoff (2004) have pointed out, the process of preparing data is like the process of refining oil. Oil starts out as a thick tarry substance mixed with impurities. It is only by going through various stages of refinement that the raw material becomes usable. Just as the most powerful engines cannot use crude oil as a fuel, the most powerful algorithms (the engines of data mining) are unlikely to find interesting patterns in unprepared data. The correctness and consistency of the data also matter the performance of data mining research. So as to boost the performance of the algorithms, the researcher tried to assess such problems which require due emphasis by applying different preprocessing techniques. Such methods are data cleaning, data

reduction, and data integration and transformation. Each of these methods will be discussed in details as follows.

4.4.1 Data Cleaning

Data Cleaning is a process which fills in missing values, removes noise and corrects data inconsistency. Usually, real world database contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2001).Consequently, data cleaning has become a must in order to improve the quality of data so as to improve the performance of the accuracy and efficiency of the data mining techniques.

This technique involves removing the records that had incomplete, noise (invalid) data and filling missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire dataset.

As a result , the researcher make use of MS-Excel 2010 built-in functions like search and replace, filtering, and auto fill mechanisms ,and WEKA to identify and fill missing value.

4.4.1.1. Missing Value Handling

Missing values refer to the values for one or more attributes in a data that do not exist. In real world application data are rarely complete. It can also be a particularly pernicious problem. Especially when the dataset is small or the number of missing fields is large, not all records with a missing field can be deleted from the sample.

Moreover, the fact that a value is missing may be significant itself. A widely applied approach is used to calculate a substitute value for missing fields, for example, the median or mean of a variable (Famili and Turney, 1997). Accordingly, the researcher has been analyzed the VCT dataset and identified missing values and take measure to solve the problem as follows. As we depicted in section 4.3.3, missing values are clearly identified and calculated the percentage of the value against total records. As a result, the entire selected attribute illustrated less than one percentage (1.00%) except campus ,

year and category attribute which of them accounts of 55.26%, 57.12%, and 64.08% respectively. To handle the problem of missing values for nominal data type attributes, replacing with modal value is recommended in Two Crows Corporation (1999).For numeric data type attribute the missing values were recommended to be replaced by the mean of that value a whole

Therefore, based on the above principles the researcher handles the missing value and WEKA preprocessing replace missing value techniques also used. WEKA fills using the most frequent (modal) value methods which is same as the above principle. Additionally, manually tracing and fixing the missed value is other technique used by researcher. Summary of handled missing value is illustrated in Table 4.22

As shown below in table 4.22 out of the selected 20 attributes 19'n of them have registered with a missing values. Accordingly, the researcher reacts to take appropriate action to clean the data.

4.4.1.2. Handling outlier value

The data stored in a database may reflect outlier – noise, exceptional case, or incomplete data object and random error in a measure of variable. These incorrect attribute values may be due to data entry problems, faulty data collection, inconsistency in naming convention or technology limitation (Han and Kamber, 2001). The authors have also explained four basic methods for handling of noise data. These are binning method, clustering, regression and combined computer and human inspection.

In this research, the researcher has identified and detected noise or outlier value from the VCT data. With the help of domain experts the identified outlier was corrected manually. As mentioned in the previous section, we tried to describe the outlier particularly in table 4.18 and 4.21 of steady and causal attribute. Accordingly, a combined effort of the researcher and domain expert were taken to identify and correct the problem of the inconsistence, incompleteness, noise or outlier.

Table 4. 22 Summary of handling missing value

Number	Attribute name and their data type	% of missing value	Replaced with	Reason/Technique applied
1	Residence :Nominal	1.64	Addis	Most frequent value
2	Category: Nominal	64.08	Community	Reason discussed under table 4.4
3	Year : Nominal	57.12	Not applicable	Reason discussed under table 4.5
4	Sex:Nominal	0.09	male	Most frequent value
5	Mar status: Nominal	0.58	Never Married	Most frequent value
6	Education :Nominal	0.99	Undergraduate	Most frequent value
7	Employment:Nominal	1.06	no	Most frequent value
8	Reas Hear:Nominal	1.59	plan	Most frequent value
9	PrevTest:Nominal	1.27	no	Most frequent value
10	EverHSex:Nominal	0.9	yes	Most frequent value
11	SusexpTime:Nominal	1.13	No experience	Most frequent value
12	Conduse:Nominal	1.15	Not Applicable	Most frequent value
13	LastCondUse:Nominal	1.09	no	Most frequent value
14	HSTI:Nominal	1.32	no	Most frequent value
15	Steady: Numric	.89	1	Most frequent value
16	Causal:Numeric	1.07	1	Most frequent value
17	SessType: Nominal	.51	Individual	Most frequent value
18	Campus:Nominal	55.26	Not applicable	Reason discussed under table 4.20
19	HIV test:Nominal	0.79	negative	Most frequent value

In table 4.20 we identified 10657("-") incomplete values the researcher used bin mean smoothing to replace the missed data as recommended by Han and Kamber (2006).

With this regards the mean of the total records is 0.95 and rounded to 1. so the replaced value will be “1”. By the same token, some value of the records exhibit inconsistency and different from the existing one which is called noise data or outlier. For instance, “1” and “01”,2 and “02”, “many” and ” too many” and other reveal inconsistence and ”-+”,” _*”,”3--,4--”,-1,-0 and much other outlier are observed from this particular attribute. In similar fashion the researcher remove 0 from prefix and keep the remaining data as a value of the record and remove many and too many value and replace with maximum value i.e. “20”. On the other hand, based on the recommendation of Han and Kamber (2006), we removed the outlier “-+”,” _*” and replace with the most frequent value (mode) which is “1” and remove suffix (- -) from “3- -“and “4- - “and keep their value as it is.

Moreover, in table 4.17, we can see same problem aforementioned above and same technique is applied to clean the data before implementation using WEKA.

4.5. Data transformation and Reduction

As Han & Kamber (2006) have stated, data mining often requires data integration or the merging of data from multiple data sources. The dataset which were used for the subsequent model building are prepared and derived from one source, AAII VCT clients records, hence no need of merging of the database.

In data transformation, the collected data are transformed or consolidated into forms appropriate for mining. The process of data transformation might include smoothing (using bin means to replace data errors), normalization, where the attribute data are scaled so as to fall within a small specified range (scaling a data inside fixed range), attribute construction, where new attributes are constructed and added from the given set of attribute to help the mining process (Chackrabarti, et al., 2009). The data needed to be reduced in order to make the analysis process manageable and cost- effective. Chackrabarti et al.(2009) and Han and kamber (2006) discussed in their books that data reduction techniques include data discretization is one of data transformation methods which is used to reduce the number of values for a given continuous

attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values, data cube aggregation, dimensional reduction (irrelevant or redundant attributes are removed), and data compression data is encoded to reduce the size, numerously reduction (models or samples are used instead of the actual data).

In this research some attributes were discretized to reduce the unlike values of the attribute to obtain knowledge (pattern) and to make the dataset suitable for data mining tools. As a result, the following section depicted the details of them.

4.5.1 Discretization and concept hierarchy generation

Han and Kamber (2006) reported that raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Discretizing the residence attribute value: this attribute consists of the nine Ethiopian regional state (Afar, Mekele, Harari , SNNP, Gambela , Benshangul, Oromiya, Amhara and Tigray) and the two City administration namely Addis and Dere. But some value is repeated with different name with same distinct value. Here it replaced with addis and the remaining distinct value is transformed starting from R1 through R9. For addis, Dere and other option as R14, R10 and 99 are used respectively.

Discretization summary of year attribute: Based on the current curriculum of tertiary education, the researcher tried to associate and classify the current academic year of user of VCT. For this reason the year category of the client will be range and transformed as Y1 to Y6 and “NY” instead of ’not applicable’ and ‘others’ are substituted.

Discretizing the category Attribute value: So as to understand with a high level of concept and easy to interpret the output of data mining, and gain advantage of it, the researcher discretize the category attribute value as follows in Table 4.23.

To decrease the size of the attribute (same value of the attribute with different values and simplicity of manipulation) the researcher tried to collocated and transform in to new value. For instance ‘staff family’ and ‘community’ almost consist of same behavioral value. So it will be redundant and should be removed.

Table 4. 23 Discretization summary of Category attribute

Attribute Name	**Old value**	**Transformed value**	Remark
category	Regular student, post graduate and student	Regular student	For the sake of easy understanding and interpretation, the most related values are grouped together.
	Extension student	Extension student	
	Summer student	Summer student	
	Staff	staff	
	Community , Staff family	Community	
	Partner	Partner	
	Other, Visitors and not applicable	Other	

Discretization summary of Education attribute: The different distinct value of educational attribute has been transformed in to high level of concept. The details has been shown in Table 4.25

Moreover, the researcher also tried to transform the distinct value of the campus or location. The old and transformed values were shown below in table 4.24

Table 4. 24 Campus attributes values discretization

Attribute Name	**Old Attribute value**	**Transformed Value**	**Remark**
Campus	Sidist Kilo	Sidist Kilo (SS)	Social science and FBE members are included
	Technology N. & S.	Technology N. & S. (Technology)	The number of Tech s.is very few , if u merged , no problem created
	Commerce	Commerce (Com)	
	Medicine	Medicine(Med)	B.lion and B.Lion , both of them belongs to same family
	Science	Science(SC)	Art and Art schools holds same data and under the constitute of science
	NA, other	Not Applicable	Hence, NA is a high level and inclusive to incorporate other.

Table 4. 25 Education Attribute values Discretization

Attribute Name	**Old Values**	**Transformed value**
Education	Illiterate	Illiterate
	Able to read	Able to read
	Primary	Primary
	secondary	secondary
	undergraduate	undergraduate
	postgraduate	postgraduate
	other, TVET, diploma and other certificate is included	other

As shown in table 4.25 the value of the attribute transformed and grouped in to high conceptual level. For the sake of analyzing the data easily using WEKA, it is also converted the transformed value in to nominal. The new grouped distinct value is"other" which transformed from "diploma","TVET" and the old name"other" educational level. Therefore, the final valid distinct values will become seven valid distinct values.

Discretizing the Reason Hear Attribute value: The primary reason the clients are visiting the centers has 22 distinct attribute values. These attribute values are grouped in to five higher knowledge level values. These are Plan, Risk, Causal, Know status and other. The details high level grouping will be illustrated below in Table 4.26

As shown in table 4.26, the researcher transformed the details old value into concrete five new values to improve the performance of the models.

Table 4. 26 Discretization of ReasHere Attribute Values

Attribute Name	**Old values**	**Transformed value**	Remark
Reason Here	"Client Risky" ,"Partner Risky", "Not trust partner" , "Other blood/fluid exposure", "Sexual assault", "Preliminary ART"	Risk	
	"Premarital" , "Marital reunion", "Family planning", "Visa applicant", "Need counseling", "Test before pregnant", "Pregnant must know", "Plan for future"	Plan	
	"TB". "death Illness of partner", "Ill/Symptoms" , "Occupational exposure",	Causal	Clients exposed to HIV causal
	"Referred" , "2nd Test (windows) " , "Confirm positive result", "Get results previous test", "Status know"	Status know	
	Other	Other	

However, from the context of HIV/AIDS, clustering this possible distinct value in to high knowledge level and removing unnecessary data help both the researcher and the machine to easily manage and generate appropriate pattern.

Accordingly the researcher by discussing the matter with domain expert grouped the value in to three distinct values. These are “No partner”, “One partner” and “Multiple partners” for both attributes. Hence, the actual values 0 and 1 are replaced by “No partner” and “One partner” respectively, whereas the value more than 1 is replaced by “Multiple partners”.

Discretizing the Values of Age Attribute: Replacing numerous values of a continuous attribute by a smaller number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining result. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts(such as numerical values for the attribute age) with high level concepts for instance Child hood, adolescent ,adult and old(Han and Kamber,2006). Accordingly, binning methods was used to scale the values of “Age” attribute. Such a meaningful discretized values used in this research are grouping the age value in to six distinct value including “0-15”, “16-20”, “21-25”, “26-35”, “36-45”, and “46+”. These groups shows pattern of “age” if any. Therefore, the actual values of the ages are manually replaced by the corresponding high level grouped values.

Table 4. 27 Discretization of age attributes value

S.N	Age attribute in range	Transformed value
1	“0-15”	A1
2	“16-20”	A2
3	“21-25”	A3
4	“26-35”	A4
5	“36-45”	A5
6	“46+”	A6

As shown in the above table range of age attribute value reduced in to six distinct high conceptual levels.

Once the data sets are pre-processed and cleaned, the next steps is develop a data mining model using WEKA 3.7.5 software which is handled in the next chapter.

CHAPTER FIVE

EXPERMINTETION AND ANALYSIS OF RESULT

In this chapter, the researcher describes the techniques that have been used in developing a model to predict HIV infection risk factor. This research incorporated the typical stages that characterize a data mining process. This study has been organized according to hybrid processing model, which is described and discussed in section 1.4.1 and 2.4.4 and shown in figure 3.3 of chapter one and two respectively. Here the researcher discuss the experimentation process by relating the steps followed ,the choice made , the task accomplished , the result obtained, evaluation of the model and results , and present it in a way that the organization can easily understand and use it.

5.1. Model Building

Modeling is one of the major tasks which is undertaken under the phase of data mining in hybrid methodology. In this phase several data mining techniques are applied and their parameters are adjusted to optimal values. Typically, different techniques can be employed for similar data mining problems. Some of the tasks include:- selecting the modeling technique, experimental setup or design, building a model and evaluating the model.

5.1.1. Selecting Modeling Technique

Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification techniques has been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, PART, J48 and SMO are used for experimentation of this study.

The researcher selected the above algorithms, easy of understanding and interpretation of the result of the model.

A PART algorithm is common types of rule induction technique which generate a model as a set of rules. In the meantime, the J48 algorithms of decision tree generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class output (Witten and Frank, 2005). Besides, SMO algorithm is available algorithm for Support vector machine implementation in WEKA to classify high dimensional data.

J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA (Two Crows Corporation,1999).It is WEKAs implementation of C4.5 top-down decision tree learners proposed by Quinlan(1986). This research used C4.5 algorithms to predict HIV infection risk factor. This algorithm is implemented by modifying parameters such as confidence factor, pruning and unpruning, changing the generalized binary split decision classification and other option available in Table 5.1. Therefore, it is very crucial to understand the available options to implement the algorithms, as it can make a significance difference in the quality of the result. In many cases, the default setting will prove adequate, but to compare results /models and attain the research objectives other opitions are considered (Witten and Frank, 2005, Han and Kamber, 2006).

Table 5. 1 Parameters for building J48 trees.

Name	Possible values	Default value	Description
M- Number of Instance	1, 2…	2	Minimum number of instances in leaves (higher values result in smaller trees).
U_ un pruned tree	yes/no	no	Use un pruned tree (the default value 'no' means that the tree is pruned).
C. confidence factor	10^{-7}- -0.5	0.25	Confidence factor used in post pruning (smaller values incur more pruning).
S-sub tree raising	yes/no	yes	Whether to consider the sub tree raising operation in post pruning.
B-use binary splits	Yes/no	no	Whether to use binary splits on nominal attributes when building the tree.

As indicated in above Table 5.1 the commonly used parameter of J48 algorithms was illustrated with various options related to tree pruning. Pruning produces fewer, more easily interpreted result. In other word, pruning can be used as a tool to correct for potential over-fitting. The algorithms recursively classifies until each leaf is pure, meaning until dataset has been categorized as close to their respective classes possible. Hence, the process ensures the maximum accuracy on the training data, but creates excessive rules.

In general, Witten and Frank (2005) explained the various option of tree pruning in J48 algorithm as follows. Basically, J48 uses two pruning methods; these are subtree replacement and subtree raising. The former method used J48 in which the nodes in a decision tree may be replaced with a leaf to reduce the number of tests along a certain path. This process starts from the leaves of the tree, and works backwards toward the root. The later methods used where a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtree raising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtree raising can be somewhat computationally complex.

Reduced-error pruning option: Error rates are used to make actual decisions about which parts of the tree to replace. There are multiple ways to do so. The simplest method is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over fitting. This approach is known as reduced-error pruning. This method reduces the overall amount of data available for training the model.

Confidence factor option: Confidence factor is the approach that seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires

a confidence threshold, which by default is set to 25 percent. This option is important for determining how specific or general the model should be. If the value of the confidence factor is lowered on the selected model, the training data is expected to fit in closely to the data we would like to test. As a result of this a more pruned or more generalized tree will be produced.

Minimum number of instances per leaf option: this is the lowest number of instances that can constitute a leaf. The higher the number the more general the tree will be. Lowering the number will produce more specific trees, as the leaves become more granular.

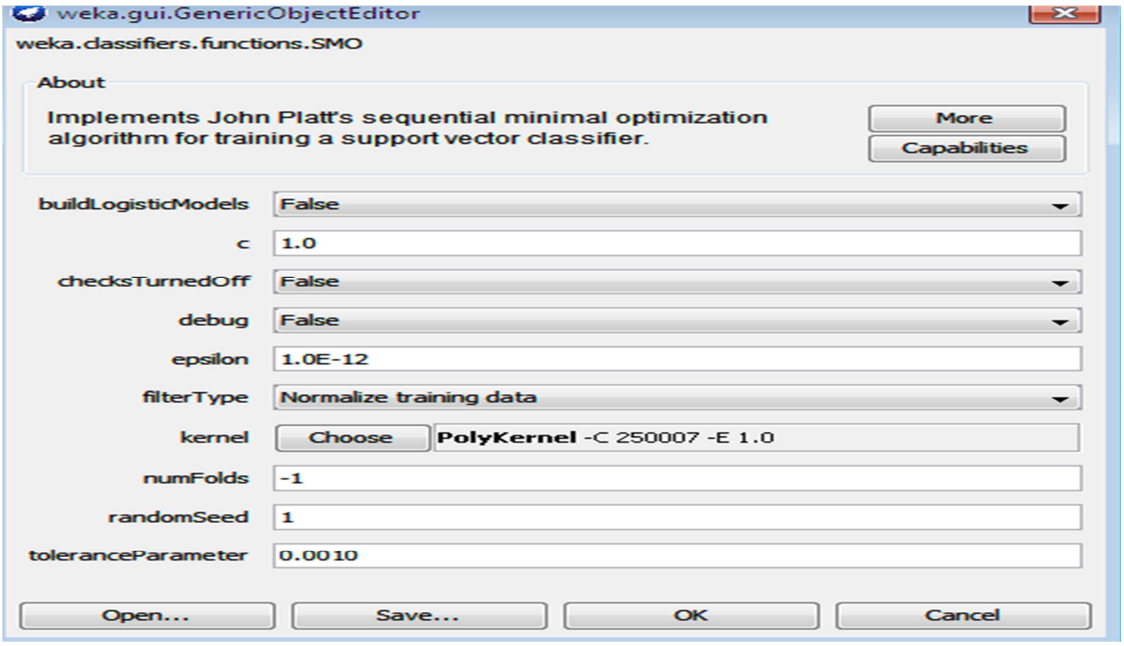
Binary split option: The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees.

Laplace smoothing option: This option is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities.

If the data mining researcher decides to employ tree pruning, it is advisable to consider the above options. Depending on how the training and test data have been defined, the performance of an unpruned tree may superficially appear better than a pruned one. This can be a result of overfitting. Hence, it is important to repeatedly experiment with models by intelligently adjusting these parameters to obtain the best set of options. According to Witten and Frank (2005), five basic and familiar parameter that usually used for data mining research is depicted in Table 5.1. PART algorithms are also separate-and-conquer rule learner proposed by (Witten and Frank, 2005). The algorithm generates sets of rules called ’decision lists ’which are ordered set of rules.

PART builds a partial C4.5 decision tree in each iteration and converts the “best” leaf into a rule. The five basic parameter mentioned for J48 algorithms decision tree are also available and applicable for PART. Since SVM popularity in 1998, it controls complexity and overfitting issues, so it works well on a wide range of practical problems (Vapnik ,1998). Due to this reason and previous studies research directions, the researcher selected to implement SVM in this study. To investigate the applicability for these research purpose SMO algorithms has used. SMO is nothing but implementation of SVM in WEKA. Before it gets to implement the software automatically convert all nominal attributes into sets of binary attributes. Then the WEKA takes the dataset as input.

Figure 5. 1 The snap shot of SMO algorithms



As shown in the above figure, the possible parameters used to implement the SVM are depicted. Besides, under kernel parameters there is also other possible parameter that is applicable for SMO to extend. The details description of the common basic parameters used in SMO are as depicted in Table 5.2. (Platt,1986).

5.1.2. Experimental Setup

In any data mining process before building a model, we need to generate a procedure or mechanism to test the model's quality and validity. For instance, in supervised data mining tasks such as classification, it is common to use classification accuracy measure or error rates as quality measures for data mining models.

Table 5. 2 Parameter for building SVM models

Name	**Default value**	**Description**
BuildLogistic Models	False	if this is true, then the output is proper probabilities rather than confidence scores
C-Complexity parameter	1.0	limits the extent to which the function is allowed to overfit the data
filterType	Normalize training data	whether you normalize the attribute values
Exponent	1.0	Used to describe the degree of the polynomial kernel
numFolds	-1	cross validation for training logistic models
Kernel	Polynkernel-c250007-E 1.0	Use to select types of kernel depending for data set(being linear and non-linear)

Besides, other standard measure including precision, recall, sensitivity and specificity are available. Therefore, the test design specifies that the dataset should be separated into training and test set, and builds the model on the training set and estimate its quality on the separate test set. With this regards Two crows Corporation (2005) reported that the process of building predictive models requires a well-defined training and validation protocol in order to insure that most accurate and robust prediction.

In this research 15,396 datasets are used for training and testing. WEKA 3.7.5 software has used to set up and measure the quality, validity and test of the selected model. For purpose of this study k-fold (10-folds) cross validation and percentage split test options are used because of its relatively low bias and variations (Witten and Frank, 2005).

Accordingly, the datasets are randomly partitioned equally into ten parts. Hence, 90% of the dataset is for training and 10 % for testing for former and the dataset are partitioned in to percentages (70-30 splits option meaning 70% of the dataset for training and remaining for testing).

In addition to these a different 3000 testing dataset is used for testing the prediction performance of the classification model developed. These testing dataset is prepared by systematic random sampling technique from the original dataset.

Witten and Frank (2005) reported that 10-fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier. To build the model of this research 19 independent and 1 dependent variables or attributes are used .The details of the selected attribute is discussed in chapter four and see the result in Appendix 2.

5.1.3. Attribute ordering

Since attribute selection is important in decision tree models, the researcher tried to rank the attribute based on information gain. It was calculated based on entropy value of the attribute. As Witten and Frank (2005) explained information gain is calculated from sum of entropy for every attribute. The formula for calculating intermediate values is:

$$\text{Info (D)} = -\sum_{i=1}^{m} Pi \log_2 Pi \qquad (5.1.)$$

Where, P_i is the probability that an arbitrary tuples in sets of training data D belongs to certain class. Info (D) is also known as the entropy of *D*. After you calculate information gain for each attribute, select the one with the highest information gain as the root node, and continue the calculation recursively until the data is completely classified by J48 algorithms in this case.

For the purpose of this research, WEKA is used to compute the information gain and rank according to the importance of the attribute for the classifier. As a result of this, the following

figure 5.2 depict ranked order of attribute based on their relevance for the reason that such attributes are very important for later experimentations by excluding the least relevant attributes.

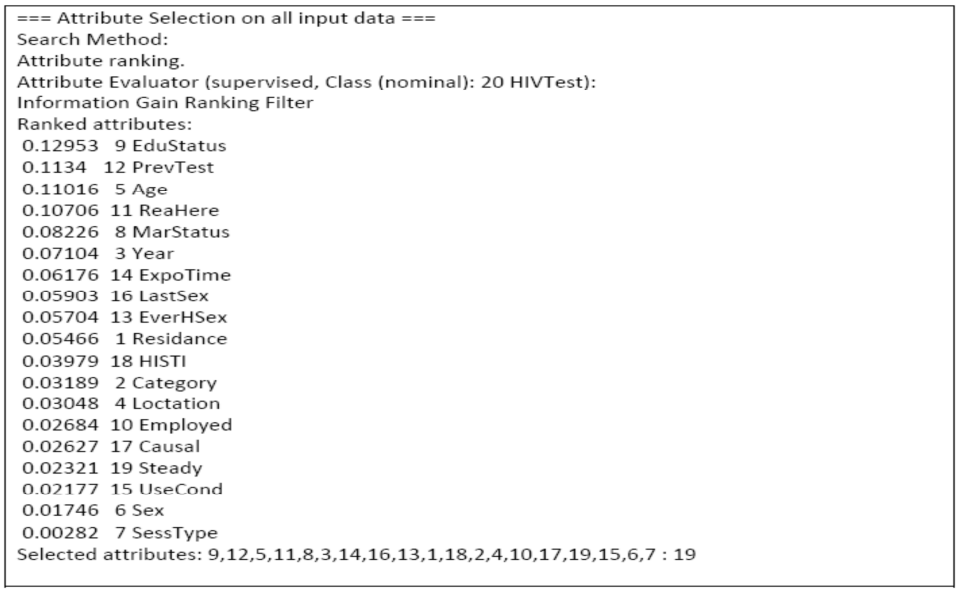


Figure 5. 2 Summary of ranked Attribute

As you can see from the result of attribute selection using entropy based information gain method of WEKA, out of 19, the top 15 the most relevant attributes are Edustatus, Previous test, Martialstat, Reasonhere, Ever had sex, Age, History of Sexual intercourse, Last sex without condom, year, residence, steady , location ,causal ,employees and exposure time.

5.1.4. Running Experiments

As it has shown in experimental setup of 5.1.2 for training and testing the classification model the researcher used two methods. **Methods One**: Percentage split method (holdout), where 70 % of the data used as training and the remaining 30 % testing. **Method Two**: K-fold cross validation methods, the data was divided into 10 folds, some fold is used as testing and the remaining folds are used as training.

Based on the above methods establishing scenario for model to be developed is very important to see the model result and analysis of each result; to compare the result of one model with the

previous one and finally help us to find out the outperforming model based on criteria of evaluation. Consequently for both of the methods following scenario has been done for each of three selected model with default parameter value of WEKA 3-7-5 software.

Scenario #1 : Pruning with all attributes

Scenario #2 : Unpruning with all attribute

Scenario #3 : Pruning with reduced attribute

Scenario #4 : Unpruning with reduced attribute

Once the modeling tool is chosen and performance evaluation criteria was established, then it followed by building model with a number of parameters that govern the model generation process. The choice of optimal parameters for the problem at hand is an iterative process, and it has to be properly explained and supported through results. So the next section present the resultant models that should be properly interpreted and their performance explained.

5.2 Model building using J48 decision tree

In Hybrid or Ciso.et al. methodology, data mining particularly model building is the crucial steps that should be deals with. Model building is an iterative process's. Therefore in this study different experiment is conducted altering the parameter with mentioned methods in section 5.1.4 using J48 algorithms, PART and SMO algorithms for building best predictive model. The parameters of each algorithm have been discussed in details in section 5.1.1.Using aforementioned methods; the experiment has been conducted with four (4) scenarios for each methods.

The researcher used 15,396 dataset with 20 attribute to investigate the model. Hence, in the real environment as everybody guess, the number of clients who have being infected with HIV is out weighted by those who have been free from the virus. This research dataset is not out of this estimation, so the actual dataset is 14,866 are HIV negative client and the remaining is positive.

However, conducting experiment with imbalanced dataset misleading classification performance and biased to the majority class (Chawla, 2003**)**. Additionally, it brings about poor decision to be taken as a result of the experiment. To coup up this and be fair the researcher used resampling technique to balance the imbalanced dataset in preprocessing phase with WEKA software. As described in chapter two of section 2.5.1.1 decision tree induction follows a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built (Han and Kamber, 2006). All of the 19 attributes, which are selected for model building, are fed as independent variables and the dependent variable of HIVTest is a target class. The algorithms builds starting at each node it will be sent either left or right according to some test. Eventually, it will reach a leaf node and be given the label associated with that leaf. In general, this research is more interested in generating rules that best predict the HIV infection and to come up with an understanding of the most important factors (variables) that can be a cause of HIV/AIDS. As mentioned in Table 5.1 J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Accordingly, based on two methods selected the following series of experiment has been conducted. Summary of experimentation with J48 algorithms result is depicted in Table 5.3

As mentioned in chapter one of methodology section 1.4.1.5, one of the compulsory step of hybrid methodology next to model building is evaluation of the model. Accordingly, the performance of the model has been evaluated based on the following criteria (see details in section1.4.1.5) including performance accuracy, confusion matrix value, and TP and FN rate, Number of leaves and size of the tree generated and ROC curves and execution time.

As shown below in Table 5.3 the experimentation has performed in two methods (in 70-30 percentage split and 10 -fold cross validation test option respectively.) The first experiment has be tested with four scenario mentioned above with 70- 30 percentage split criteria. When we compare the result of method I. Scenario # 2(unpruning with all attribute) is the best model based on correctly classified instance(out of 4619 , 4459 instances are correctly classified), performance register (95.7 % accuracy and TP rate of 96 % (classifying correctly HIV negative

client as HIV negative and HIV positive as positive). Besides, in terms of minimum number of leaves and size of tree (both indicate the complexity of generated tree) and time taken to build a models are Scenario # 4 (pruning with reduced attribute) and scenario # 3 (unpruing with reduced attribute) is best model respectively.

Table 5. 3 Experimentation result of J48 Algorithms with two methods

Model Characteristics	Experiment (Scenario)							
	1	2	**3**	4	1	**2**	3	4
Test option	**I**				**II**			
pruned	yes	no	**yes**	no	**Yes**	no	Yes	no
Attribute	All	All	**reduced**	reduced	**ALL**	ALL	reduced	reduced
Accuracy	**95.2**	**95.7**	**93.8**	**94.5**	**95.5**	**96.3**	**94.9**	**95.6**
Time taken	1.1	0.9	**0.88**	0.45	**1.5**	0.64	1.0	0.48
#of leaves	1707	2074	**1614**	2021	**1707**	2074	1614	2021
Size of trees	2106	2554	**1966**	2459	**2106**	2554	1966	2449
AV .TPR(%)	95	96	**94**	95	**95**	96	95	96
AV.FPR(%)	5	4	**6**	5	**5**	4	5	4
AV.PR	0.96	0.97	**0.95**	0.95	**0.97**	0.97	0.95	0.96
AV.RR	0.95	0.97	**0.95**	0.95	**0.97**	0.97	0.95	0.96
AV.ROC	0.97	0.98	**0.97**	0.98	**0.98**	0.98	0.98	0.98
CCI	4397	4459	**4331**	4364	**14706**	14833	14619	14722
ICI	222	160	**288**	255	**790**	563	777	674

Key: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate, I: 70-30 percentage split option (Holdout), II: 10- fold cross validation

The last compares in method one is made between Average ROC curve rate which registered a performance of 97 and 98 % in pruned and unpruned respectively irrespective of the number attribute. Above all from the method I, even though the complexity of the generated tree is high, correctly classifying instance and performance of the model is better .As a result Scenario # 2 (Building decision tree unpruning with all attribute) is selected.

Similarly, in the second experiment or method two (10-fold cross validation cases) four of the stated trial was conducted. When we compare the result of each scenario developed model, scenario # 2 (building decision tree unpruned with all attribute) is best based on accuracy which registered 96.3 % and correctly classified instance which is accounts 14,833 out of 15396 . Furthermore, based on criteria of minimum time taken for building model, scenario # 4 (building model with unprunned with reduced attribute) is best.

Analogously, the number of tree and size of tree for scenario # 3 is minimums so it is best because it reduced the complexity of the generated tree. The average ROC curve performance measure indicates same for all scenario and account of 98%. Finally when we compare the actual correctly classified instance based on their labeled class i.e. classifying HIV negative client as HIV negative for all scenario, irrespective of number of attribute unpruned register better than pruned. By the same token, from methods two, still scenario number two better than other model in terms of accuracy, correctly classified instance and ROC curve.

However, when we compare the size and leaves of tree of unpruned J48 model, the number is enormous and complex relative to pruned one. As a result of this the algorithms might not reach optimality and generate more generalized decision tree rules. Han and Kamber (2005) also reported this is because of overfitting problem. This is a fundamental problem in learning of algorithms. Besides, such situation has its own impact in classification performance particularly classifying unseen or new instance. Subsequently to solve the problem the researcher selected pruned scenario that perform better accuracy. Accordingly, scenario #1 (Building pruned

decision tree with all attribute) of 10- fold cross validation(method II) selected as the best J48 decision tree model. See the details confusion matrix in appendix 3.1

5.2.1. Comparison of J48 decision tree model

In general when we compare both methods in terms of classification model accuracy 10 fold cross validation (method II) is better than 70-30 percentage split .The accuracy ranges from 94.9 to 96.3 but 70-30 percentage split 93.8 to 95.5. Additionally, WEKA experimenter tool was used to compare of the J48 model developed with various scenarios. The result of the comparison is depicted in Figure 5.3

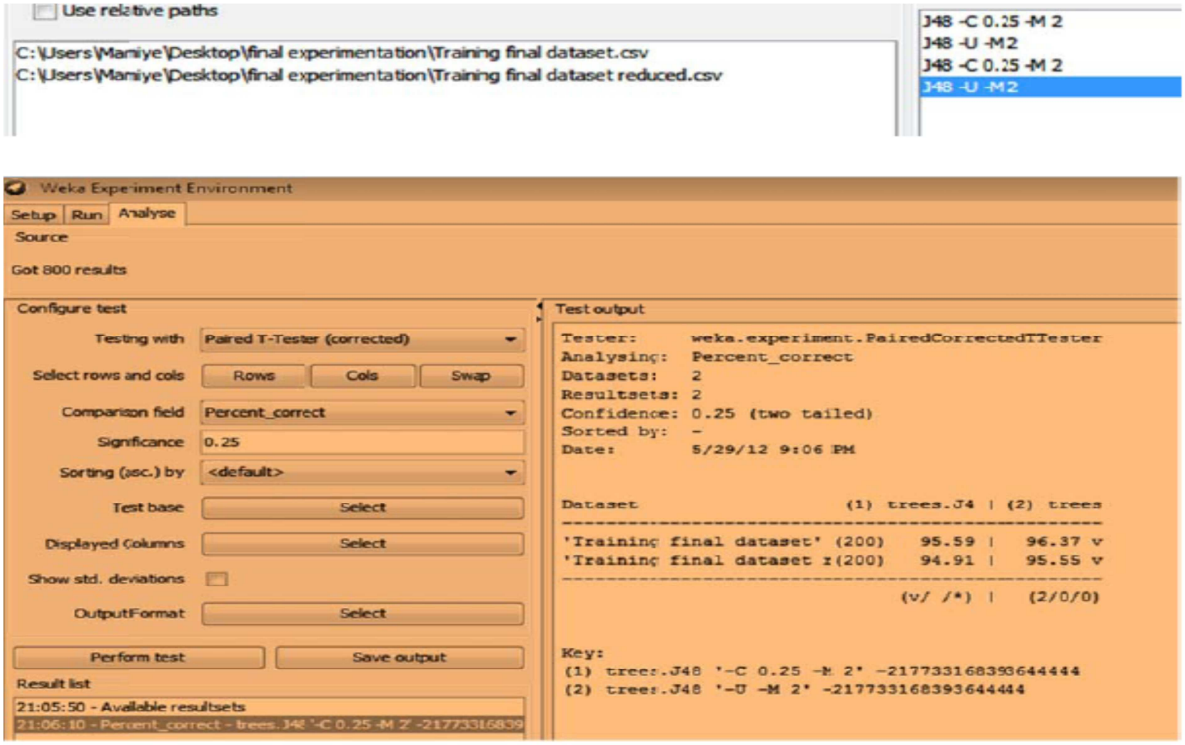


Figure 5. 3 Experimental dialog box with compered J48 algorithms

Consequently, as shown in the above figure and Table 5.3 experimentation, irrespective of reduced or all attribute the performance of unpruned J48 algorithms register better than pruned.

However, by the same token aforementioned, the researcher selected scenario # 1 (building decision tree pruned with all attribute) of 10 fold cross validation which registered **95.6**% accuracy as a final J48 decision tree model for training and testing the dataset and validated the performance of the model after getting reduced its complexity with increasing minimum instance parameters.

The detailed confusion matrix as a result of pruned J48 algorithm with all attribute has been discussed below in Table 5.4 and shown in Appendix 3.1

Table 5. 4 Summary of confusion matrix for pruned J48 decision tree with all attribute

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	7040	636	7676
	Positive	54	7666	7720
Total		7094	8302	15396

The base of this research is determining the risk factor of HIV infection and finding pattern that shows the status of the HIV. In this respect, the J48 algorithms classify how those instances are correctly and incorrectly classified in labeled class.

So as indicated in the above Table 5.4 based on the 10–fold cross validation test option the J48 learning algorithm scored an accuracy of 95.5%. This result shows that out of the total training datasets 14706 (95.5 %.) records are correctly classified, while only 690 (4.5%) of the records are incorrectly classified.

Besides, out of the total 7676 actually HIV negative clients, only 7040 (91.7%) clients are classified as HIV negatives and the rest are misclassified as HIV positive. And out of the total 7720 actually HIV positive clients, only 7666 (99.3%) clients are classified as HIV positive and

the rest are misclassified as HIV negative. This means the model has better performance in terms of correctly classifying HIV positives than HIV negatives.

According to Shermer (2002) report hypothesis testing, error can be categorized in to two based on the classification of being true as false and false as true. These errors are called Type I(false postive error) and Type II(False negative error) So in this research context of the experiment, classifying HIV negatives as HIV positive (Type II error) may lead the client anxiety and incurs cost of VCT testing. Whereas classifying HIV positives as HIV negative (Type I error) may have also its own risk such as spread of the disease. Hence, Type I error is more serious than Type II error in this case since it can affect more people.

Furthermore, evaluating the model based on sensitivity and specificity are very significance for decision making. For that reason .the result of the above confusion matrix indicate that the sensitivity of this test was (7666/7720) = 99.3 % and the specificity was (7040/7676) = 91.7%. The test indicates that the models appear to be pretty good. Because, based on the evaluation criteria, the classifier correctly classify clients as HIV negative who has actually diseases free with 91.7% accuracy and classify clients as HIV positive who actually has the deasese with 99.3%. However, the false positive and false negative rate which corresponds to classification error and register performance accuracy of 0.95 and 0.05 respectively.

As discussed in Table 5.3, the size of the tree and the number of leaves produced from method two of scenario # 1 model training was 2106 and 1707 respectively. This seems that it is difficult to navigate through all the nodes of the tree in order to derive out with valid sets of rule. Therefore, to make ease the process of generating rule sets or to make it more understandable, the researcher attempted to modify the default values of the parameters so as to minimize the size of the tree and number of leaves.

In this regards, the minNumObj (minimum number of instances in a leaf) parameter was tried with 10,20,30,50 100 and 200. However, the minNumObj set to 200 gives a better tree size and

minimal accuracy compared with the other trials. With this value of the minNumObj the process of classifying records proceeds until the number of records at each leaf reached 200. The result of the experiment depicted in Table 5.5 and the instant of the confusion matrix with minNumObj= 200 shown in appendix 3.2

Table 5. 5 Summary of pruned J48 algorithms with minNumObj = 200

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	6033	1643	7676
	Positive	1797	5923	7720
Total		7830	7566	15396

The experiment has shown an improvement in the number of leaves and tree size. The size of the tree is dropped from 2106 to 95 and the number of the leaves decreased from 1707 to 76. As we can see from Table 5.4, the resulting confusion matrix shows that the J48 decision tree algorithm scored 77.7 % accuracy. That means out of the total 15396 records 11956 (77.7 %) are correctly classified and the remaining are misclassified.

In other words, the confusion matrix of this experiment has shown that 6033 (78.6%) of the 7676 total HIV negative clients are correctly classified as disease free. But 1643 (22.4 %) of them are misclassified as HIV positive class. Besides the result shows that out of 7720 total HIV positive clients records 5923 (76.7 %) of the records are correctly classified while 1797 (23.3 %) of the records are misclassified as HIV negatives being HIV positives.When we compare this experiment result, classifying clients who has HIV positives as HIV positives outweigh from those classifying clients as HIV negative. Overall, nevertheless execution time reduce from 1.5 to 0.52 seconds ,and the tree size and the number of the leaves decreased significantly as shown above, the accuracy of the J48 decision tree algorithm in this experiment is poorer than the first

experiment with the default parameter value of the minNumObj. Besides, the accuracy of the model lowered from 95.5 to 77.7 % as minNumObj getting large which shown a considerable change. Hence the first selected algorithms is optimal and exhaustively see the whole thing irrespective of is efficiency. And can conclude that from this experiment minNumobj and accuracy of model performance have shown inversely related pattern. Additionally to keep the performance of the model the first experiment with the default minNumObj parameter value is better and taken as the best J48 decision tree model. The generated possible decision tree is shown in Appendix 4.

As described in section 5.1.2 randomly selected datasets are used to validate the performance of the model. Accordingly, once it has been developed the J48 model using the whole training dataset with 10 –fold cross validation test option and registered the possible capacity of the model that accurately classify new instance. However, validating the developed model with independent dataset will enhance believability and applicability of the model. To do so the researcher, using supplied test option re-evaluate the model by mentioned dataset. The result of the reevaluation of the model depicted as follows in Table 5.6 and sees the snapshot of the model in appendix 3.3

Table 5. 6 Result of re-evaluation of J48 decision tree with 3000 dataset

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	1409	103	1512
	Positive	73	1415	1488
Total		1482	1518	3000

The performance of the model with this testing dataset was 94.1%. This result shows that out of the 3000 testing datasets, the developed decision tree classification model predicted around 2789 (94.1%) records correctly.

In general, the researcher exhaustively discussed based on the result of the model with domain expert to determine variable that cause HIV infection. According to the experiment depicted in Table 5.4, the J48 model used only eleven attribute out of 19 attributes including PrevTest, EverHsex,Year,Age,Usecond,ReasHear, causal, Category, ExpoTime, Sex , and employed. So from this model, these are the determinate attribute of HIV infection. However, the researcher cross checked the validity based on expert opinion. According to Health expert, marstatus and edustatus attribute are also can be used as the cause of HIV infection.

Hence, the above mentioned attribute plus experts opinions (overall 13 attributes) are the most determinat factor of the HIV infection. Besides, the pattern /rules obtained by setting pruned, confidence factor 0.25 and test option 10 fold cross validation J48 decision tree with all attribute (in this case 19) model is useful for identifying HIV infection risk factor and pay attention on those important clients behavior by health professional and by the university management in particular.

5.3 Model building using PART Rule Induction Algorithms

The second data mining classification technique applied in this research was PART Rule induction algorithms. As mentioned in chapter two of literature review section 2.5.1.3,there are many rule induction algorithms but the researcher selected PART for the reason that PART has the ability and potential to produce accurate and easily interpretable patterns/ rules that helps to achieve the research objectives. PART is a separate-and-conquer rule learner like that of decision tree and proposed by Witten and Frank(2005).The algorithm is apply an iterative process and produces set of “decision lists” which is ordered set of rules. It works by generating a rule that covers a subset of the training examples and then removing all examples covered by the rules from the training set. This process is repeated until there are no examples covered left to

cover. The final rule set is the collection of the rules discovered at every iterations of the process. The rules are in standard form of IF-THEN rules.

To build the Rule induction model, WEKA software package and the same 15,396 VCT dataset was used as an input respectively. The experiment will be performed analogously as we did in former model. That means the experiment was divided in two methods as methods one: experiment with 70-30 percentage split criteria for four scenario and Method two as with 10-fold cross validation evaluate the four scenario again. The parameters are partially adjusted and default value was used with reduced and all attribute. Accordingly, the experiment of all scenarios with methods one will be illustrated below in Table 5.7

Table 5. 7 Experiment result of PART algorithms for two methods

Model Characteristics	Experiment (Scenario)							
	1	2	**3**	4	1	**2**	3	4
Test option	**I**				**II**			
pruned	yes	no	**yes**	no	Yes	**no**	Yes	no
Attribute	All	All	**reduced**	reduced	ALL	**ALL**	reduced	reduced
Accuracy	**95.7**	96.5	**94.7**	95.1	**96.7**	**97.3**	95.4	96.2
Time taken	**23.7**	37.3	**15.2**	24.7	24.0	**28.8**	15.5	24.5
No. of Rules	374	508	393	524	374	**508**	393	524
AV .TPR	0.96	0.97	**0.95**	0.95	0.97	**0.97**	0.95	0.96
AV.FPR	4	3	**5**	5	3	**3**	5	4
AV.PR	0.96	0.97	**0.95**	0.95	0.97	**0.97**	0.96	0.96
AV.RR	0.96	0.97	**0.95**	0.95	0.97	**0.97**	0.95	0.96
AV.ROC	0.97	0.98	**0.97**	0.97	0.98	**0.98**	0.98	0.98
CCI	4421	4459	**4373**	4392	14886	**14975**	14687	14814
ICI	198	160	**246**	227	510	**421**	709	582

Key: CCI: Correctly classified Instance, **ICI** (Incorrectly classified Instance), **Accuracy**: Registered performance of model, **AV**: Average, **TPR**: True Positive Rate. **FPR**: False Positives

Rate, **ROC**: Relative Optical character curve, PR: precision rate, **RR:** Recall rate, **I**: 70-30 percentage split option (Holdout), **II**: 10- fold cross validation

As shown in the above table, the registered performance is better that induction rule learner in case of unpruned instead of pruned one. But by same token aforementioned in case of J48 decision tree classifier, the researcher selected the one which register best with pruned case irrespective of the number of attribute and methods.

Consequently, among the four scenarios experimented with 70-30 percentage split (method I) with varying parameters, scenario #1 registered better performance of 95.7%. This shows that out of the training set of 4619 records 4421 (95.7%) of the records are correctly classified, while 198 (4.3%) of the records are misclassified.

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	2145	195	2340
	Positive	3	2276	2279
Total		2148	2471	4619

Table 5. 8 Confusion matrix of PART algorithm with 70-30 percentage-split

Furthermore, the resulting confusion matrix also shown in Table 5.8 that out of the total 2340 HIV negative clients 2145 (91.7%) of them are correctly classified in their respective class, while 195 (8.3%) of the records are incorrectly classified in the HIV positive classes or wrongly classified client as HIV positive without having really the diseases). In other hand out of the total number of HIV positive clients 2276(99.86%) of them are correctly classified as HIV infection as positive but 3 (0.14%) of the client records are misclassified.

Generally, from the above experiment for the two classes, classifying clients as HIV positive outweigh than HIV negative clients. Even though, this model performs better than ever developed so far PART algorithms, it is performing smaller than the one performed by decision tree J48 algorithms in case of method one.

The second method employed in this research to experiments with 10 fold cross validation test options. The result of the four scenario experiment was displayed in Table 5.7.

In the same way, the experiment was made with different scenario for same test option and registered better performance in case of scenario #1 which accounts 96.78% of accuracy than other model. In this method of experiment, the PART algorithm brings better performance in case of the size of tree or rules than in unpruned ones. Because, it reduce the complexity and increase easily understanding of the rules. Additionally, even better performance has been registered than J48 decision tree.

As aforementioned above in J48 decision tree model building, reducing attribute with different test option doesn't bring significance change in the performance of the model rather remains same or very little difference that of all attribute scenarios. This indicates that the reduced attribute (such as SessType, location, causal and sex) has very little or no role in determining HIV infection. The same is true in rule induction learner except minor change in case of scenario with unpruned parameter irrespective of number of attribute. In other hand , those attribute that are selected as the best model has been considered as the major determinate of the HIN infection. From Table 5.6 of method two scenario # 1(PART rule learner with 10 fold cross validation with all attribute) register an accuracy of 96.7% and depicted the predicted result in table 5.8 and see the details confusion matrix in Appendix 3.4.

As shown in table 5.9 below the test result of the confusion matrix the model developed by the PART Algorithm with the 10- fold cross validation, the model scored an accuracy of **96.7**%. This shows that from the total 15396 test data, 14866 (**96.78**%. of the records are correctly classified,

while 530 (3.3%) of them are misclassified. In addition the above table also shows that from the total 7676 HIV negatives clients records 7182 (93.5%) of them are correctly classified as HIV negatives, while 494 (6.5%) of the records are misclassified as HIV positives.

Table 5. 9 Final selected best PART algorithms using 10- fold cross test option

		Predicted HIV infection test result		
		Negative	Positive	Total
HIV infection status (Actual)	Negative	7182	494	7676
	Positive	16	7704	7726
	Total	7198	8198	15396
	10 fold cross validation test option: PART			

Additionally, you can also observe that out of the total 7726 actual HIV positives clients records 7704 (99.79%) of the records are correctly classified as HIV positives clients, while 16(1.21%) of them are incorrectly classified in the HIV negative class. when we compare the HIV positive and negatives class, classifying clients as HIV positive correctly perform better than HIV negative.

To sum up, in the above experiment similarly as of J48 decision tree classifier (method II or 10-fold cross validation) that is conducted using the PART algorithms with all attribute and default parameters values generates a better classification model with a better classification accuracy than 70-30 percentage split test options. So again we selected scenario #1 of 10.fold cross validation model as the best model of PART for classifying clients records as HIV negative and positive and in the mean time we can determine variable that could be risk factor of HIV infection see the snap shot of confusion matrix in appendix 3.4

5.3.1. Feature selection for PART algorithms

One of methods to improve the performance of classification accuracy is implementing feature or attribute selection. In line with J48 algorithms by default has the system automatically select best attribute and generate decision tree only based on best selected feature. For this research pupose the algorithms determines the listed variable mentioned above as the determinate attribute of HIV infection. But PART has no such facility to select the best attribute automatically instead the researcher using default parameter, tried to select feature. The result of this feature includes Residance, Year, Age, MarStatus, EduStatus, ReaHere, PrevTest, EverHSex and instance class are selected out of 20 attribute.

When we compared the performance of the PART algorithms selected as best model running without feature selection and with featured selected running model the former registered 96.7% and the latter 87.2% of accuracy respectively which is a big change or reduction in performance (9.5%). The latter model confusion matrix (**see appendix 3.5**) developed by the PART Algorithm with the 10- fold cross validation, the model scored an accuracy of 87.2%. This shows that from the total 15396 test data, 13426 (87.2.%). of the records are correctly classified, while 1970 (12.8%) of them are misclassified. Additionally, from the total 7676 HIV negatives clients records 6501 (84..7%) of them are correctly classified as HIV negatives, while 1175 (15.3%) of the records are misclassified as HIV positives. And out of total 7720 HIV positive clients records 6925 (89..7%) of them are correctly classified as HIV positive, while 795 (11.3%) of the records are misclassified as HIV negatives..

According to the domain expert evaluation of the selected features are not the only but there is also other determinant variable such as usecodom,employed, steady and HISTI that could be risk factor of HIV infection. As a result the researcher and the domain expert selected the first model (PART algorithms with 10 fold cross validation and all attribute) as the best model that used for predicting HIV infection.

5.3.1. Analyzing Interesting Rules from PART algorithms

It is shown that the best performing model among all models developed using PART was scenario # 1 in Method II illustrated in Table 5.6.The experiment conducted with 20 attribute including instance class, default parameter value (confidence value=0.25,MinNumObj=2 unpruned= false).Running the PART rule induction algorithms on supplied dataset generates rules in plain text. The number of rules generates are 374 and registered accuracy of 96.7%.

To reduce the complexity of the rule and easy understandability the decision rule generated by the above model, the researcher tried by varing the minNumobj(values of 0.2 ,0.25,0.30, 0.40 ,0.50) and minNumObj (value of 2,5,10) are experimented by combining the paired parameter. As a result PART algorithms with parameter confidence factor 0.5 and minNumObj =2 register (96.8%) accuracy which is better than other the selected best PART model. However, the number of rules generated still lengthy so we rely on the first choice.

The partial PART decision list was shown in appendix 5**.** This model generates 121 rules among this some of the most interesting pattern or rules discovered will be illustrates as follows.

1. IF PrevTest = “no” AND EduStatus =” Illiterate “AND LastSex = “no” AND Employed =” no” AND Age = “20-25” then Then the status of client will be HIV positive . Forty five records of the client correctly satisfy this rule. (45.0)
2. IF PrevTest = “no” AND Residence = “Addis” AND Steady =”One partner” AND Year = “3rd year student” AND Category = “staff parent” AND sex = “Female” Then the status of client will be HIV positive. Twelve records of the client correctly satisfay this rule.
3. IF the clients visited voluntary counselling and testing center because of risky and without having sex and whose gender is female then the probability of the client HIV infection is Positive. These rules are correctly found in 34 of the records.
4. IF PrevTest = “yes+ve” AND Employed =” NO” AND UseCond =” Never” AND Age = “20-25” then the status of HIV infection is Positives. These rules are correctly found in 21 of the records.

5. IF Residance = “addis” AND PrevTest = “ yes-ve “ AND Causal = “one partner” AND MarStatus = ” separated” AND HISTI =” No“ AND ReaHere =” plan” AND Employed = “Yes “ AND Age = “20-25” then the status of Client HIV infection is Postives. This rules is correctely found in 18 records.

6. If Residance = “Addis” AND MarStatus = “not married” AND UseCond = “ never”AND LastSex = “ used condom” AND Sex = ”female” AND Age = “20-23”then the status of clent HIV infection is Positive (21.0).

7. PrevTest = “No “AND ReaHere =” causal” AND Year = “3rd year student” AND UseCond= “ never”AND

8. HISTI =” No “AND Age = “” greater than 45 then the status of the client HIV infection is Positive (111.0/3.0)

9. IF Residance = “Addis” AND Loctation = “sadist killo” AND MarStatus = “Married” AND EverHSex = “Yes “AND UseCond = “Never” AND LastSex = “No”AND EduStatus = “other “ ANDReaHere = “risk “AND Age = “>= 46” then Pos (33.0/1.0)

10. If Residance = “addis” AND LastSex =” No” AND EduStatus = ”undergraduate” AND Causal = “one partner” AND Category = “summer student” then client becomes HIV positive (19.0)

11. PrevTest = “yes-ve “ AND Residence = “addis” AND UseCond = “never” AND Causal = “one partner” AND ReaHere =” risk “AND Loctation = “sidist kilo” AND ExpoTime = ” 4to6month” then client most likely HIV positive (14.0)

12. PrevTest = “yes-ve “ AND Residence = “addis” AND UseCond = “never” AND Causal = “one partner” AND ReaHere =” risk “AND Loctation = “sidist kilo” AND ExpoTime = ” 4to6month” then client most likely HIV positive (14.0)

13. IF PrevTest = “No” AND Residance = “Addis” AND Age = “greater than 46 ” AND HISTI = “No “ AND ExpoTime = “less1m “ AND MarStatus =” Marriage” then HIV Positive (22.0/2.0)

14. IF PrevTest = “no” AND ReaHere = “plan” AND EduStatus = “secondary” AND Age = “20.25” Then the status of Clients HIV infection will be negatives. (712.0/4.0)

15. IF Residence = “Addis” AND Steady =”One partner” AND Year = “3rd year student” AND Category = “others” AND Causal =”One partner” AND Age = “greater than or equals to 46” Then the status of client will be HIV positive.(13/5)

16. IF Residence = “Addis” AND Steady = “One partner” AND Year = “3rd year student AND Category = “others” AND MarStatus = “never married “AND Sex = “female Then the probability of Clients HIV Infection status is Positive.(8/1)

17. IF Residence = “Addis” AND Category = “Regular student” AND PrevTest = “No” AND MarStatus = “separated “AND Employed = “yes” Then the probability of the Clients HIV Infection status is Positive.(6/2)

18. IF Residence = “Addis” AND Category = “Regular student” AND MarStatus = “divorced “AND ExpoTime = “no experience” Then the probability of the Clients HIV Infection status is Positive.(11/3)

19. IF Residence = “Addis” AND Category = “Regular student “AND PrevTest = “yes “HIV +ve” AND Sex = “Female” Then the probability of the Clients HIV Infection status is Positive.(10/1)

20. IF Residence = “Addis” AND Category = “Regular student “AND PrevTest = “No” AND EduStatus = “secondary” AND UseCond = “never” Then the probability of the Clients HIV Infection status is Positive.(3/1)

As observed from the above partial list of rules, the classifier has used all attribute to construct rules and provided the class predicted by the model unlike J48 algorithms generated rules. The numeric values which appeared in the bracket next to the class label indicates the number of correctly and incorrectly classified records respectively.

For instance Rule 18 interpreted as clients who came from or live in Addis and registered in regular program and his marital status is divorced and she/he has no experience to exposed to HIV negative then the classifier grouped the behavior under HIV infection status of postive(11/13.0). Meaning in the dataset 1470 records that are exactly satisfied this rule and 17 records indicate misclassified to these rules. Similarly, the remaining rules can be interpreted in the above way.

5.4. Model Building using SMO Algorithms

The third classification task used in this research was Support vector machine. As mentioned in chapter two, the best SVM algorithms that available in WEKA software is sequential minimum optimization algorithms. It works based on risk minimization principles.

To build the model ,even though SMO handles missing value and other noise data automatically the researcher initially clean and handle such issue in chapter four of section data preparation and data cleaning. The same dataset (15396) are used as input for SMO.

The same methods were employed for experimentation as used in the above two models. These are the 10-fold cross validation and the percentage split with 70-30 test option with reduced and all attribute and default values for training and testing the model were used.

Method I: The first experiment of the SVM model building is performed using the SMO algorithm with the default value of percentage split criteria of 70-30 test option. Table 5.10 shows the resulting of two methods and table 5.11 indicate the confusion matrix of the model developed using this algorithm for method one.

Table 5. 10 experimental result of SMO Algorithms with both methods

Experiment	**Scheme**	**# of attribute**	**Accuracy (%)**	**Time (Sec.)**	**CCI**	**ICI**	AVTPR (%	AVFPR (%)	AVROC(%)
Method I: Experiment result with 70:30 percentage Split option									
Scenario # 1	SMO-C1.0-L..	**20**	80.7	2689	3730	889	81	19	81
Scenario #2	SMO -C 1.0…	**16**	80.6	2667	3724	895	81	19	81
Method II: Experiment varying parameter with 10- fold test option									
Scenario #1	SMO -C 1.0….	20	**81**	**2638**	12475	**2921**	**81**	19	81
Scenario # 2	SMO -C 1.0…..	16	80.6	**5463**	12410	2986	97	19	81

As shown in the table 5.10 in method one, we can only observe minor difference in time taken to build a model and remaining experimental result almost similar irrespective of the number of attributes used for each scenario. In this regards, reducing attribute shows only reduce model building time without changing anything else.

So we can generalize that the reduced attribute as mentioned in previous algorithms has no significance for being the risk factor of the HIV infection because as indicate in both method no change at all was made for each model performance. Let me see the confusion matrix of method one scenario # 1.

Table 5. 11 Confusion matrix of SMO with 70-30 percentage split option

		Predicted HIV infection test result		
		Negative	Positive	Total
HIV infection status (Actual)	Negative	1914	426	2340
	Positive	463	1816	2279
	Total	2377	2242	4619
	70-30 percentage split test option: SMO			

As can be seen from the confusion matrix that resulted from the model developed by the SMO algorithms with the 70-30 percentage split, the model scored an accuracy of 80.7%. This shows that from the total of 2340 HIV negatives 1914 (81.7%) of the records are correctly classified as HIV negatives, while 426 (18.3%) of them are misclassified as HIV positives.

Furthermore when we see (appendix 3. 6a) the confusion matrix of the remaining scenario in method two, similarly scenario # 1(Running SMO algorithms with 10 fold cross validation) scored an accuracy of 81 %. This shows that from the total of 15,396 dataset 12475 records are correctly classified as per records of the original dataset and the remaining are misclassified.

The confusion matrix showed in appendix 3.6b out of 7676 total HIV negatives 6318(82,3%) of the records are correctly classified as HIV negatives while 1358 (18.3%) of them are misclassified as HIV positives. In other hand, out of the total 7720 records of HIV positive, 6151(79.7%) are correctly classified as HIV positive and remaining are misclassified as negative.

As one observe from the above Table 5.10 methods II of scenario # 1 is selected the best smo. Meaning the experiment that is conducted with 10-fold cross validation, all attribute and default parameter register best accuracy of 81%.

5.5. Comparison of J48, PART and SMO models

Selecting a better classification technique for building a model, which performs best in handling the prediction and identifying risk factor of HIV infection is one of the aims of this study. For that reason, the three selected classification model with respective best performance accuracy are listed in table 5.12

Table 5. 12 comparison of the result of the selected three models

Types of algorithms	Performance registered (%)	Time taken(sec.)	Correctly classified	Misclassified
J48	**95.5**	1.5	14706	790
PART	96.6	3.37	14866	530
SMO	81	2637	12475	2921
10 fold cross validation test option				

The result showed in table 5.12 that PART rule induction outperforms J48 decision tree and SMO by 1.2% and 15.7% in identifying determinant risk factor of HIV infection. The reason for the PART rule induction to perform better than other is because of linearity of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular HIV infection.

Moreover, in terms of ease and simplicity to the user the PART rule induction is more self-explanatory because it produces in forms of "If then". It generates rules that can be presented in simple human language.

Therefore, it is reasonable to conclude that the PART is more appropriate to this particular case than the J48 decision tree and SMO. So, the model that is developed with the PART rule induction classification technique is taken as the final working classification model to determine or predict HIV infection risk factor to assist both the organization and the university management in case of decision making in matter of HIV related issue.

5.6. Evaluation of Discovered Knowledge

Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006).Besides, data preparation task is a crucial so that data mining tools can understand the data.Because without doing so the accuracy of the discovered knowledge will be poor.

Thus, data cleaning and analysis methods and outlier mining methods becomes a must to tackle such gab before discovered knowledge so as to enhance the efficiency and effectiveness of the data mining technique. In this effort, the researcher tried to assess such problems which require due emphasis while making the data ready for applying data integration and other different data reduction techniques. See the details effort that has been done in this regards in chapter four.

As discussed in chapter two and five, in selection of model building, three classification models was selected and developed for reason explained below. Firstly, the J48 decision tree algorithm

is chosen for is simplicity easy interpretability of the result generated by it. Secondly PART is selected similarly as of J48 and simplicity and generated list is easy to interpret and associate the problem domain. Finally, because of the recommendation of previous research studied in area of HIV/AIDS the researcher selected SMO algorithms. Above all, three of the model was experimented, evaluated and compared as a result with 10 fold cross validation test option , PART rule induction learner was used as a final model for study.

From 374 all possible list of decision rules generated only few of the rules are discussed. These rules are unambiguous, easy to interpret and understand. see more about PART decision rule set on Appendix 5.

As illustrated in appendix 5, all attributes are used for construction of decision rule set. As indicated in section 5.3.2 experiment has been implemented by applying WEKA feature selection technique. However the performance is less than those implemented with all attribute .As a result of this research output is evaluated based on all attribute generated decision rule set. Hence, the most interesting rules or discovered knowledge were:

Rule 1: IF Residence = “Addis”AND Category = “Regular student” AND MarStatus = “divorced “AND ExpoTime = “no experience” then HIV Positive.(11/3)

Rule 2: IF ReaHere=”risk” AND ExpoTime =”no experience” AND Sex=”Female”: Postives(34.0)

Rule 3: IF PrevTest = “yes-ve “ AND Residence = “addis” AND UseCond = “never” AND Causal = “one partner” AND ReaHere =” risk “AND Loctation = “sidist kilo” AND ExpoTime = ” 4to6month” then client most likely HIV positive (14.0)

Rule 4: IF PrevTest = “no” AND EduStatus =” Illiterate “AND LastSex = “no” AND Employed =” no” AND Age = “20-25”: postives (45.0).

Rule 5: IF PrevTest = “yes+ve” AND Employed =” NO” AND UseCond =” Never” AND Age = “20-25”: Postive(21.0)

Rule 6: IF Residance = “addis” AND PrevTest = “ yes-ve “ AND Causal = “one partner” AND MarStatus = ” separated” AND HISTI =” No“ AND ReaHere =” plan” AND Employed = “Yes “ AND Age = “20-25” : Positive(18)

Rule 7: PrevTest = “No “AND ReaHere =” causal” AND Year = “3rd year student” AND UseCond= “ never”AND HISTI =” No “AND Age = “greater than 46” : positive(111.0/3.0)

Rule 8: PrevTest = “No “AND ReaHere =” causal” AND Year = “3rd year student” AND UseCond= “ never”AND HISTI =” No “AND Age = “” greater than 45 then the status of the client HIV infection is Positive (111.0/3.0)

Rule 9: IF Residance = “Addis” AND Loctation = “sidist kilo” AND MarStatus = “Married” AND EverHSex = “Yes “AND UseCond = “Never” AND LastSex = “No”AND EduStatus = “other “ AND ReaHere = “risk “AND Age = “>= 46” then Pos (33.0/1.0):

Rule 10: IF PrevTest = “no” AND ReaHere = “plan” AND EduStatus = “secondary” AND Age = “20.25” Then the status of Clients HIV infection will be negatives. (712.0/4.0)

Rule11.If PrevTest=”yes HIV +ve” AND ReaHear=”other”AND ExpoTime=“donot known” AND MarStat= “never married” LastSex = “yes”Then HIV negative.

Rule 12 IF Residence = “Addis” AND Category = “Regular student “AND PrevTest = “yes “HIV +ve” AND Sex = “Female”: Positive(10/1).

The rules that are accessible above indicate the possible conditions in which VCT service clients record could be classified in each of HIV Negative and positives classes. All attributes are participated in construction of decision rule set but as observed above the contribution in each of the experiment and domain expert validation four of the attribute (Sesstype, sex, location and steady) are less determinate of HIV infection.

From the above list of partial PART rule induction generated decision rule set or discovered rules/ knowledge, we can observe how the relation of attribute contribute towards HIV infection.

A client more likely to be HIV negative even if previous test is HIV positive. Rule 11 is a good indicator of this fact. The domain expertise will also agree with this finding because it happens because of malfunction of the laboratory or some man made problems. Other hand the laboratory result of the client may be HIV positives even if previous laboratory test of a client said HIV negative. This because the virus is not visible within window period (meaning a client can confirm his test result after three month). Rule 4, 7 and 8 shows this relationship. The domain expert also supported this idea by saying clinically approved situation.

Moreover clients whose previous test was HIV positive (Rule 5) might not be change for second time laboratory result. This is for the reason that clients have been infected and cannot recover from diseases rather prolong their life span by taking “tablet” for limited year. Furthermore, Clients whose previous HIV status is negative have also been shown as negative during repeat testing (Rule 10). This may reflect that there is some behavioral change after getting VCT service. UNFPA (2002) reported that understanding status of HIV negative can serve as a strong motivating factor to remain negative and is considered as evidence from the existing knowledge in the domain.

According to Rule 7 and 9 age group above 45 years old are more exposed to HIV infection. For instance, Rule 9 that client who comes from Addis and currently found in sadist kilo campus whose age was greater than 45. Additionally who got married and never use condom and their educational status is unmentioned; they came in VCT center because of something that they consider risk for HIV infection then the client is suspected to HIV infection. Based the classification accuracy of the PART algorithms 33 of the record dataset is implied this. In this regard the domain expert reported that this situation is true sometimes it occurred because, more attention is given through different media on productive age (age from 20-30). Health professional do not always test older people for HIV/AIDS and so may miss some cases during routine check-ups. Additionally, sometimes people above 50 often mistake signs of HIV/AIDS for the aches and pains of normal aging, so they are less likely than younger people to get tested for the disease.

Furthermore, this research finding indicates that those people whose ages between 20-25 are more exposed to the HIV infection. For instance, Rule 4, 5 and 6 reported the effect of this rule. To elaborate clients who never used condom and their age is among 20 and 25 irrespective of employment status may be more susceptible to HIV infection. The USAID (2006) report has also stated that since AIDS deaths are concentrated in the 25 to 45 age group, communities with high

rates of HIV infections lose disproportionate numbers of parents and experienced workers and create gaps that are difficult for society to fill.

In real environment and many research suggest, females are more susceptibility to HIV Infection than male (UNAIDS, 2000). This is indicated by Rule 2 and 12. Rule 12 report that those clients whose residence was Addis and registered for regular program having job and knows her status as HIV positive earlier would be classified as HIV positive. Moreover, that female client that has no experience to susceptible to to HIV infection and consult VCT center for reason of risky condition is vulnerable to HIV infection.

Lastly, this research finding showed that clients whose level of education or year in which they achieved as 3rd year university student are exposed to HIV infection. Rule 7 and 8 reported this situation. The domain expert explained this condition is true. Additionally, in real environment of university freshman students are more exposable than senior student. This is because most of these students are new to university and exposed to various thing such as chewing chat, drinking and smoking cigarette. This is the major factor of HIV infection in university (Negatu and Seman, 2011). By doing so, they can be easily exposed to the diseases.

In general, the domain expertise also agree with the general idea that the model produced and adds some fact that some clients also vulnerable as a result of addiction which has not included as a behavior of clients in dataset. So having knowledge of their addiction of the client is very important to determine and predict their status.

The finding interpreted above age, previous test, cause, exposure time, marital status, last time use of condom in sexual intercourse, use of condom, sex, educational status, employment status, residence, category, year and reason for taking VCT services are the major determinant factor of HIV infections of clients.Some of the attributes (Location and sesstype) that the algorithms assumed less important also less important for domain expert in determining HIV infection status of this study.

The classification model used in this research to predict as risk factor of HIV infections are pretty good .They come up with reasonable accuracy performance. Among this model PART rule learner register better than other and an accuracy of 96.7 %.

The overall experimentation process of experimentation has been performed following the diagram.

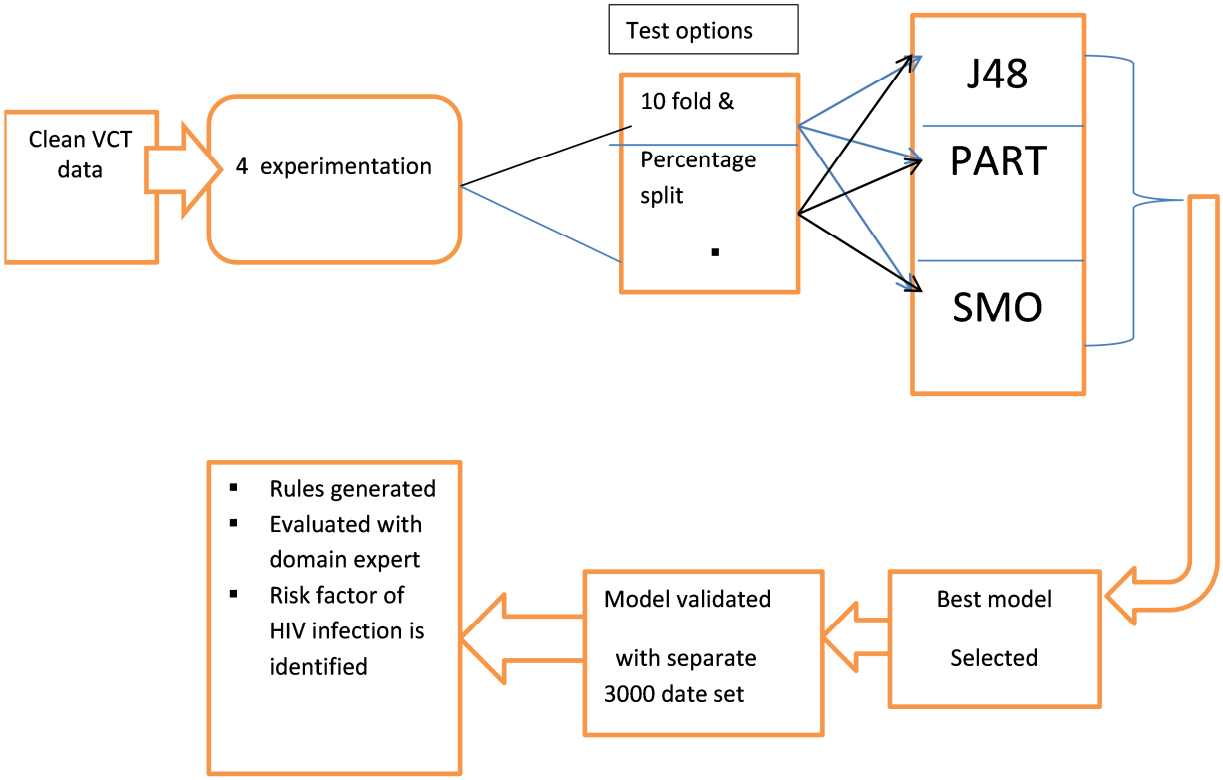


Figure 5.4 Graphical representation of Experimentation performed in this thesis

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

A major challenge facing healthcare organizations (VCT Service Provider, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable (Srinivas et al., 2010).Besides, health care centers including VCT service provider must also minimize the cost of clinical tests.

To this end, they can achieve these results by employing appropriate computer-based information and/or decision support systems particularly application of data mining technology. Health care related data including VCT data is massive. Hence, such data must be analyzed by organization using data mining technology. Hence application of data mining techniques may help in answering several important and critical questions by extraction of useful knowledge that enables support for cost-savings and decision making.

In this research, an attempt has been made to apply the data mining technology in support of identifying and predicting HIV infection risk factor using VCT data. Data mining technology basically follows an iterative process consists of: Data collection, Data preparation and understanding, model building and evaluation. As discussed different data mining methodology in chapter two , there is a possibility of additional steps to above list for instance KDD includes nine steps. The iterative nature of the process assist the data miner to back and forth at different and fix it where problem arise.

Accordingly, this research used and followed the six steps Cios et al (2007) methodology to investigate the stated problem. The methodology has strictly been followed while undertaking the experimentation. The data used in this research has been gathered from African AIDS

Initiatives International VCT center. Once the data has been collected (15396 dataset), it has been preprocessed and prepared in a format suitable for the DM tasks. Data preparation took considerable time of the study.

The study was conducted using three classification techniques namely decision tree, rule induction and support vector machine. For model building and experimentation J48, PART and SMO algorithms are used. Experimentation was conducted using four scenarios in two methods for each algorithm except for SMO. For SMO only two scenarios were used for the two methods with default value and the output of this is used as a comparison for former two algorithms.

By changing the training test options and the default parameter values of the algorithm, these models are tested and evaluated. As a result PART rule induction algorithms registered batter performance with 96.7 % accuracy running with 10-fold cross validation with default parameter using 20 attribute than any experimentation done for this research purpose. The performance of the model is validated using randomly selected 3000 datasets and registered performance accuracy of 95.8% which is pretty good model.

According to the result of PART algorithms, all the 20 attributes are used to construct the decision rule set. Besides, using WEKA feature attribute selection is tried and tested the performance is 77 %. This implies that decrease in performance resulted because of missing of relevant attribute. To this end, with extensive discussion of domain expert and “educated guesses” out of the 20 attribute, only the following attributes are selected as the determinate variables of HIV infections has been identified.

These are previous test of HIV, use of condom, Reason hear in VCT center, History of sexual Intercourse, employment status, HIV infection causal situation, Last sex with condom, exposure time for risk situation (number of partner the client faced), marital status, educational status, year, age .

Besides that studying the behavior of the university community related to addiction such as chewing chat, smoking cigarette and drinking are important to identify the risk factor of HIV infection.

Moreover, the finding of this research indicates that females are more vulnerable than men to HIV/AIDS. Third year university students are more susceptible to HIV infection. Previous test of client who have HIV negative and for second time test is negative. The laboratory result may be may be positive because the virus may not be visible until end of three month. A client who has got HIV positive may be result in second test "negative "which is because of health professional or malfunctioning of laboratory test or equipment.

Analogously, J48 and SMO algorithms were also register an encouraging performance of classification accuracy. Both model correctly classify new instance of VCT clients as 95.5 % and 81 % with 10 fold cross validation with all attribute and default parameter. When we compare the result of three model developed based on accuracy registered to correctly classify new instance of VCT data, PART perform better.

In general, in this study only few experiments are carried out using J48, PART and SMO with few parameter setting. Missing values are simply replaced by most frequent values using WEKA 3.7.5 and manually for few attribute in consultation of domain expert which may sometimes create bias so that if time allows it is better to use other mechanism like filling extensively by consultation of domain expert. Furthermore, the researcher faced challenging in experimentation of support vector machine because of time consuming, balancing of the unbalanced dataset and getting of VCT dataset.

In spite of the challenges and weaknesses in data pre-processing and limited application of algorithms the results are encouraging. The results obtained from this research indicate that data mining is useful in bringing relevant information to the service providers as well as decision makers.

6.2. Recommendation

In this research work an attempt has been made to find out the potential applicability of data mining technology to support the prevalence of HIV activity at Addis Ababa University community and surrounding based on the data accumulated on VCT service by AAII. Even though, this research has been done for the academic exercise its results are found promising to be applied in addressing practical problems in prevention and control of HIV /AIDS.

Hence, based on the findings of this research, the following recommendations are forwarded:

- This research has proven the applicability of PART algorithms which automatically discover hidden knowledge that are interesting and accepted by domain expert. However, this research output doesn't integrate with knowledge based system. To do so, the researcher recommends implementing the discovered classification rules with domain knowledge as a knowledge based system that could be helpful for health professional, epidemiologists and other researchers who are conducting HIV/AIDS care, control and prevention.

- Attribute such as session type and campus were not good discriminate variables for predicting HIV infection, hence to investigate further research in case of prevalence of HIV in higher education institute, these attributes will not be recommended

- This research has attempted to determine HIV infection risk factor using classification data mining task only. The association that exists between the selected attributes related to HIV infection was not considered. Therefore, research can be conducted to see the degree of association that coexists between those attributes and HIV infection is forwarded.

- For this work, balanced dataset is used to determine the risk factor of HIV infection. However, researches can also be conducted using actual imbalanced data and other data mining technique such as Class Confidence Proportion Decision Tree (CCPDT). This is a robust decision tree algorithm for imbalanced datasets.
- This research has been attempted to determine the risk factor of HIV infection with limited data set and 20 attributes collected from Addis Ababa university community and its surroundings. Further researches can also be conducted by increasing the number of attributes and datasets from all higher education institutions in Ethiopia.
- The research also attempted to test the applicability of data mining in identifying HIV infection risk factor. In doing so, implementing using more data mining techniques such as Naïve Bayes and Multilayer perception may boost the performance of the model and helps to compare the result with this research finding.

REFERENCESE

AAU (2006).HIV/AIDS Policy for Addis Ababa University (AAU).Prepared under the joint Auspices of the Addis Ababa University and African AIDS Initiatives International Inc. Addis Ababa,Ethiopia:AAU press

AAU to Host 60th Graduation Day (AAUa). Press release July 2011 .Rretrieved on March 2012 from www.aau.edu.et

Abdinasir A, Gurja B, Sentayehu T (2002). Knowledge, Attitude, Practice and Behavior (KAPB): Survey on HIV/AIDS among AAU Students, AAU, Ethiopia: UNFPA, *ISAPSO*.

Abdul-Kareem,S., Baba.S. , Zubairi, Y. Z., Wahid, M. I. A. (2000). Artificial Neural network (ANN) as a tool for medical prognosis. Sage publication

Abraham Tesso (2005) Application of Data mining Technology To identify the Determinate risk factors of HIV infection and To find Their association Rules: A Case of Center for Disease control and Prevention. Master Thesis. Addis Ababa University: School of Information Science.

Addis Ababa Highlights (AAH) (2006) magazine issue vol.3.(2)Retrieved March 2012 from www.unep.org/roa/Addis_Ababa_Site/.../February2006.pdf

Africa AIDS Initiative International (AAII) (2006). HIV/AIDS, STDS and Reproductive Health: Knowledge, Attitude, Practice and Behavior (KAPB) survey on AAU Students, Final Survey Report Addis Ababa, unpublished.

Agrawal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rule in large databases', San Jose, California: IBM Almaden Research Center

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996) Fast discovery of association rules: In Advances in Knowledge Discovery and Data Mining, pages 307–328. AAAI Press

AIDS (n.d.) All about AIDS Discovery. Retrieved On march 2012 URL: http://library.thinkquest.org/03oct/01335/en/txt/discovery.html

Behavioral Surveillance Survey (BSS) Ethiopia (2005). Round Two. AddisAbaba, Ethiopia : MOH/HAPCO, AAU, CSA, EPHA

Berry, M. & Linoff , G. (2004). Data mining techniques for marketing, sales, and customer relationship management. (2nd ed.).Indiana: Wiley publishing.

……. (2000) Data Mining Techniques for Marketing Sales and Customer Support. New York: John Wiley & Sons.

Berry, Michael J.A. and Gordon Linoff(1997). Data Mining Techniques For Marketing Sales and Customer Support. New York: John Wiley & Sons.

Berson ,Alex, Smith ,Stephen, and Thearling ,Kurt(n.d.) An Overview of Data Mining Techniques Retrive on Feb, 2012.From http://www.thearling.com/text/ dmtechniques/dmtechniques.htm

Beyene P, Solomon B, Yared M (1997). AIDS and college Students in Addis Ababa: A Study of Knowledge, Attitude and Behavior. Addis Ababa ,Ethiopia. : J. Health Dev. 11(2):20-30.

Bharti, K Jain, S and Shukla, S (2010) Fuzzy K-mean Clustering Via J48 for Intrusiion Detection System.International Journal of Computer Science and Information,1(4).

Birru Asmare(2009). Application of Data mining Technology to support VCT for HIV: A Case of Center for Disease control and Prevention. Master Thesis .Addis Ababa University. School of Information Science.

Cabena, P., Stadler,R., Hadjinian ,p. , Zanasi, A., and Verhees,J.(1998). Discovering data mining from concept to implementation. New Jersey: Prentice Hall

Chakrabarti ,s, Cox,E, Frank ,E., Hartmut ,G.R., Han, J. Jiang,X. Kamber,M and, Witten, Ian H.(2009)Data mining: know it all. Burlington. San Francisco, CA, USA: Morga Kaufmann Publishers

Chapman ,P., Clinton, J., Kerber ,R., Khabaza, T., Daimlerchrysler ,T. R., and Shearer ,C(.2000) "CRISP-DM 1.0: Step-by-step data mining guide." CRISP-DM consortium. SPSS.USA

Chawla, Nitesh V.(2003)." C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure". Workshop paper on Learning from Imbalanced Datasets II, Washington DC: ICML,

Cios, K ,Witold, P, Roman, S and Kurgan ,A. (2007). Data Mining: A Knowledge Discovery Approach, New York, USA: Springer,

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. Machine Learning,.3(4):261-283.Avilable at www.citeulike.org/user/janch/article/9304405.

Daniele M. and Satheesh R(n.d) . Improving tax administration with data mining, Elite Analytics, LLC. Retrived march2012 from http://www.spss.ch/file.php?file=/upload/1123070263_Improving%20tax%20administration

Deshpande, S P, and V M Thakare.(2010). Data mining system and applications : a review. International Journal 1 (1): 32-44.

DMSHO (2011).Data mining for Successful Healthcare Organizations (DMSHO)Retrived March 2012 from http://www.shamsgroup.com/Portals/0/Articles_in_the_media/DataMining.

Dunham, M. H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

Eapen,A.G.(2004) Application of Data mining in Medical Applications.Master Thesis . University of Waterloo. Waterloo, Ontario, Canada.Retrived on April 2012from

Elias Lemuye(2011).HIV Status Predictive Modeling Using Data Mining Technology. Master Thesis .Addis Ababa University.School of Information Science and Public Health. Unpublished

Elias Worku (2009). Internet –Based Technology as source of sexual and HIV/AIDS Related Health Information among AAU students. Unpublished Thesis

Famili, A and Turney, P.(1997). Data preprocessing and Intelligent Data analysis. Institute of Information Technology, National research council, Canada

Family Health International (FHI) (2004).HIV Voluntary Counseling and Testing: A Reference Guide for Counselors and Trainers. Retrived on March 2012 from URL: www.fhi360.org/en/HIVAIDS/pub/guide/vcttoolkitref.htm

Fayyad, Usma, Piatetsky-shpiro, G. and smyth, padharic.(1996). From Data Mining to knowledge Discovery in Databases. Available URL: http://citeseer.nj.nec.com/fayyad96from.html

FDRE HIV/AIDS PC (2010). FDRE HIV/AIDS Prevention and Control Office. Report on progress towards implementation of the UN Declaration of Commitment on HIV/AIDS Retrived On March 2010. From: http://www.unaids.org/en/dataanalysis/monitoring countryprogress/2010progressreportssubmittedbycountries/ethiopia_2010_country_progress_report_en.pdf , Accessed on December,2011

Frank,E, and Witten ,Ian .(1998).Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning,144-151,

Garbus L (2002) HIV/AIDS in Ethiopia, Country AIDS Policy Analysis Project, San Francisco: AIDS Policy Research Center, University of California

Getinet T. (2009). Self-reported sexual experiences, sexual conduct and safer sex practices of AAU undergraduate male and female student's on the context of HIV/AIDS pandemics, A PhD dissertation Buffalo, USA: the State University of New York

GHIVR (2011) Progress report 2011: Global HIV/AIDS Response(GHIVR). World Health Organization HIV/AIDS Department, Switzerland, Geneva.

GHSSH (2011).Global Health Sector Strategy on HIV/AIDS(GHSSH)2011–2015. Geneva,World HealthOrganization, 2011 Retrived on March 2012 from (http://whqlibdoc.who.int/publications/2011/9789241501651_eng.pdf, accessed March 2012).

Government of Ethiopia GoE (2004) A Comprehensive Strategic Plan to Combat HIV/AIDS Epidemic in Ethiopia (2004 – 2007), Final Report, Addis Ababa

Grzymala-busse, jerzy w.(n.d.) Rule Induction. University of Kansas Retrieved on February,2012.From people.eecs.ku.edu/~jerzy/c95-perugia.pdf

Gupta, s., Kumar, d. and Sharma, A.(2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis . Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 (2).

Hailom, B., Aklilu, K. , Sara, B. and ,Kate, S.(2005).The System-Wide Effects of the Global Fund in Ethiopia Baseline Study Report. Maryland,Abt Associates Inc. Retrieved from URL: www.theglobalfund.org/.../library/Library_IEPNADF194_Report_en.March 2012

Han, J & Kamber, M. (2001).Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

---------- (2006).Data mining: concepts and techniques. (2nd ed.). San Francisco: Morgan Kaufmann Pubisher

Hipp, J. and Ulrich, H. (2000) 'Algorithms for Association Rule Mining- A General Survey and Comparison', Tubingen: University of Tubingen

UNAIDSHIV) (2000).HIV testing methods (UNAIDSHIV): UNAIDS Technical Update. Geneva, UNAIDS. Retrieved March 2012 From URL: http://data.unaids.org/Publications/irc-pub01/jc379-vct_en.pdf

Hogg, D. W., Jackson, C., ˙Zytkow, A. N., Irwin, M., Webster, R., and Tremaine,H(2006). Rates of disease progression according to initial HAART regimen: A collaborative analysis of 12 prospective cohort studies. Journal of Infectious Diseases 194: 612-22

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning. 11:63-91.

Jinhong, L Bingru, Y and Wei, S (2009). A New Data Mining Process Model for Aluminum Electrolysis. Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA'09) Qingdao, P. R. China : Academy Publisher. Pp.193-195.

Kantardzic ,Mehmed(2003). Data mining: concepts, models, methods, and algorithms. The University of California; Wiley-Inter science

Kathy, L. and Stacey.(2009). From the Ground Up: User Involvement in the Development of an Information System for use in HIV Voluntary Counselling and Testing. Retrieved on March 2012fromURL:http://cit.mak.ac.ug/iccir/downloads/ICCIR_09/Kathy%20Lynch%20and%20Stacey%20Lynch_09.pdf

Koh, Hian C. and Tan, Gerald (2005). Data mining applications in Healthcare. Journal of Health Information Management,19(2) Available at: http://www.himss.org/content/files/Code%20109_Data%20mining%20in%20healthcare_JHIM_.pdf accessed on April 2012

Kohavi ,Ron(1995). The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189,

Kurgan and Musilek (2006) A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, Vol. 21:1, 1–24. 2006, Cambridge University Press

Larvac, Nada.(1998) Data Mining in Medicine : Selected Techniques and Applications. Retrived On April 2012 from URL.: http://citeseer.nj.nec.com/lavrac98data.html

Lee,W., Stolfo, S.J., and Mok, K.W. (1999). A Data Mining Framework for Building Intrusion Detection Models. In IEEE Symposium on Security and Privacy (pp. 120–132).

LLGP (2007). Lessons Learned and Good Practices on SRH and HIV/AIDS Prevention: Summary of Project Performances of 12 Local NGOs Supported by UNFPA/NORAD. Unpublished.

Lloyd -Williams, Michael. (1997).Discovering the Hidden secrets in your Data the data Mining approach to Information. Retrieved March 2012 from URL:http://informationr.net/ir/3-2/paper36.html

Madigan ,Elizabeth, Curet, O. and Zrinyi, Miklos (2008). Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns.Human Resource for Health licensee BioMed Central Ltd.,6(2).1-6

Mamcenko, Jelena and Beleviciute. Inga (2007). Data Mining for Knowledge Management in Technology Enhanced Learning. Journal of Knowledge Management: P.115-119.

Matsatsinis ,N. F., Ioannidou, E. ,and Grigoroudis ,E..(n.d.)Customer satisfaction using data mining techniques Chania. . Technical University of Crete.

Ministry of Health (MoH)(2000).Health and Health Related Indicators, Planning and Programming Department, MoH.unpublished

Ministry of Health (MOHHA, 2006) - HAPCO. Accelerated access to HIV/AIDS prevention, care and treatment inEthiopia. Road map 2007-2008/10. Accessed on March 2012 from URL: http://fitun.etharc.org/resources/healthpolicyandmanagementresources/roadmap2007-2010ethiopia.pdf

Ministry of Health(MOH) (2006).AIDS in Ethiopia. Fifth Report. Disease Prevention and Conttol Department, MOH.

Moshkovich,Helen M., Mechitov,Alexander I., and Olson,David L. (2002) .Rule induction in data mining: effect of ordinal scales. Expert systems with applications. Elsevier Science Ltd. . Technical University of Crete. 22 p.303-311,

Mung'ala, L., Taegtmeyer ,M., Kilonzo, N., Morgan, G. and Theobald ,S.(1995).A community-based counselling service as a potential outlet for condom distribution. 9th International Conference of AIDS and STD in Africa. Kampala, Uganda.

My Data Mining Weblog (MDMW).(2008).Rule learner (or Rule Induction) Retrived Febrary 2012 From: http://mydatamine.com/?tag=induction-algorithms

National AIDS Council (NAC)(2001)Strategic Framework for the National Response to HIV/AIDS in Ethiopia. (2001-2005).

Nigatu and Seman (2011).Attitudes and practices on HIV preventions among students of higher education institutions in Ethiopia: The case of Addis Ababa University.Available at http://interesjournals.org/ER/pdf/2011/February/Nigatu%20and%20Seman.pdf accessed October,2011

Obenshain, Mary K.(2004). Application of data mining techniques to healthcare data.: Infection control and hospital epidemiology .The official journal of the Society of Hospital Epidemiologists of America 25 (8): 690-5.

Pal, N.R., Jain, L.C., (2005).Advanced Techniques in Knowledge Discovery and Data Mining, Verlag:Springer

Palaniappan and Awang (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. IJCSNS International Journal of Computer Science and Network Security,8 (8)

Pete, C. (2000).Step by step Data mining guide. CRISP-DM consortium,.6: 60-67

Piatetsky-Shapiro, Gregory.(2000) .Knowledge Discovery in Databases: 10 Years After. SIGKDDExplorations.1Online.Internet. Retrived on march 2012 http://www.kdnuggets.com/gpspubs/sigkddexplorations-kdd-10-years.html

Plate, Bert,J, Grace,J, and Band(1997).A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer. Advances in Neural Information processing :MIT press retrived on April 2012 from http://citeseer.nj.nec.com/plate96comparison.html

Platt, John (1998).Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. MIT Press.

Popa, H., Pop, D., Negru, V. & Zaharie, D. (2007). AgentDiscover: A multi-agent system for knowledge discovery from databases. In Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing.

Prather, M. Dentener, F. , Derwent, R. , Dlugokencky, E. , Holland, E. , Isaksen, I. , Katima, J. ,Kirchhoff, V. and Matson, P.(2001)Medical Data Mining : Knowledge Discovery in a clinical Data Woehouse. Retrived on April 2012 fromURL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput605/2008/pdf/medical_data_mining_prather.pdf Accessed on October,2011

WKDD (2008) . Advanced Knowledge Discovery and Data Mining. IEEE discovery and Data Mining, Workshop on Knowledge Discovery and Data Mining. University of Adelaide.; IEEE Computer Society.

Quinlane , J.R. (1986). Induction of Decision Trees. Kluwer Academic publisher. Boston (1), 81-106

Raghavan, P. , Manning, C.D. , Schütze, H(2002) .Data Mining Trends and Issues. Retrived On April 2012 from: http://citeseer.nj.nec.com/138316.html

Razak,A. and Bakar,A. (2006). Association Rules Mining in Asthma Patients Profile Dataset. Chiang Mai Journal of Science, 33(1): 23-33 (ISSN: 0125-2526).

Rea, Alan (1995). Data mining an introduction Version 2.0, Queens University Belfast.)].261– 283.

Refaat, Mamdouh (2007).Data Preparation for Data Mining Using SAS. sanfransisco: Mamdouh Refaat publisher

Roiger, Richard. And Geatz, Michael (2003). Data mining: a tutorial-based primer.Addison Wesley,

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C 35 (4): 476–487.

Santos , Manuel Filipe,& Azevedo,Ana (2008).KDD, SEMMA AND CRISP-DM: a parallel overview .. IADIS European Confrence data miing.p.182-187.

Schneider H et al(2010). Urban–rural inequalities in access to ART: results from facility based surveys in South Africa. AIDS 2010 – XVIII International AIDS Conference, Vienna, 18–23 Retrieved March 2012 from URL: http://www.iasociety.org/Default.aspz?pageid=12&abstracted=200738997).

Sheikh, Khalid (2008). Using Decision Trees to predict customer behavior. Indian Expressed group. Retrieved on May 2012 http://expresscomputeronline.com/20040412/technology01.shtml

Shermer, Michael (2002).The Skeptic Encyclopedia of Pseudoscience 2 volume set.ABC-CLIO., California

Silltow, John(2006). Data mining: Tools and Technique. Retrived on March 2012

From at http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/

Smolander, K., Tahvanainen, V.P., Lyytinen, K., (1990) How to Combine Tools and Methods in Practise - a Field Study. In: Lecture Notes in Computer Science, Second Nordic Conference CAiSE'90, Stockholm, Sweden. pp. 195-211.

Srinivas, K. et al. (2010) Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSEI) Vol. 02, No. 02, 250-255

Tefera B, Challi J, and Yoseph M (2004). KAP about HIV/AIDS, VCT Among students of Jimma University, Jimma Zone, SW, Ethiopia. Ethiop. J. Health Sci. Vol.14, Specal issue.

Teka T (1993). College Students' Attitude and knowledge of AIDS. Ethiop. Med. J. 314:233-237.

Tekelu(2010). Application of Data mining Technology on Antiretroviral Therapy(ART) data: a case of Adama and Assella Hospital. Master Thesis .Addis Ababa University. School of Information Science

Trybula, Walter J.(1997) Data Mining and Knowledge Discovery. ed. Williams, Martha E. Annual Review Information Science and Technology., Vol. 32, Information Today, Inc. New Jersey, USA: Medford,

Two crows Corporation (1999), Introduction to data mining and knowledge discovery (2nd Ed).by Edelstein, H., A. Potomac, MD: Two Crows Corp.

…… (2005) Introduction to data mining and knowledge discovery (3rd Ed).by Edelstein, H., A. Potomac, MD: Two Crows Corp.

USAID (2006). Bringing information to decision makers for global effectiveness:how HIV and

AIDS affect population. Population Reference Bureau: Washington.

UNAIDS (2010). Global report: UNAIDS report on the global AIDS epidemic. Vol.2007, issue, Geneva.

UNAIDS(2011).World AIDS Day Report of 2011. Geneva, UNAIDS.Retrived From: www.unaids.org/.../unaids/.../unaidspublication/2011/JC2216 Accessed on March 2012.

UNAIDS/WHO (2010). AIDS epidemic update UNAIDS 20 avenues Appia CH-1211 Geneva 27 Switzerland.

UNAIDSa(2011). The Global HIV/AIDS Epidemic: Fact Sheet, 2011. Geneva, UNAIDS. Retrieved from: www.kff.org/hivaids/upload/3030-16.pdf .Accessed March, 2012.

UNAIDSb(2010). Report on the Global AIDS Epidemic; Geneva, UNAIDS.Retrieved On March 2012 URL: www.unaids.org/globalreport/global_report.htm

UNAIDSc(2011). AIDS at 30: Nations at the crossroads; 2011.Retrieved On March 2012 URL: www.unaids.org/unaids_resources/aidsat30/aids-at-30.pdf

USAIDHPI (USAID Health Policy Initiative). Equity: quantify inequalities in access to health services and health status. Washington, DC, Futures Group, Health Policy Initiative, Task Order 1, 2010 (http://www.healthpolicyinitiative.com/Publications/Documents/1274_1_ EQUITY_Quantify_FINAL_Sept_2010_acc.pdf, accessed March 2012).

UNFPA (2002).Voluntary counseling and testing (VCT) for HIV Prevention. HIV Prevention Now Program Brief 5: 1-2.

Vararuk, A., Petrounias , I. , & Kodogiannis, V.(2008).Data mining techniques for HIV/AIDS data management in Thailand. Journal of Enterprise Information Management,21(1):52-70.

Witten, I. H., & Frank, E. ,2000 ,"Data mining concepts" ,New York, Morgan-Kaufmann

Wong, M. L., Leung, K. S.(2000). Data Mining using Grammar Based Genetic Programming and application. Springer

World Health Organization (WHOa).(2010) Towards Universal Access: Scaling up priority HIV/AIDS interventions in the health sector. Progress Report, September 2010. Retrieved March 2012 http://www.who.int/hiv/pub/2010progressreport/summary_en.pdf

World Health Organization (WHOb) (2004). Policy statement on HIV testing. Retrieved on March, 2012 URL:http://www.who.int/hiv/pub/vct/statement/en/index.html

World Health Organization (WHOS) (2005).Scaling-up HIV testing and counselling services: a toolkit for program managers. Geneva, Switzerland

Appendix 1: Standard Format Of AAII VCTUser Registration Form

Appendix 2: Descriptions of Selected Attributes

No.	Attributes	Descriptions	Codes & Corresponding Values
1	Age	Age of the clients	{A1=0-15, A2=15-19, A3=20-25,A4 =26-35, A5=36-45,A6 >= 46}
2	Sex	Sex of the clients	{Male=M, Female=F}
3	SessType	Session type of the clients	{ indv=Individual , Coup=Couple , other=Group}
4	MarStat	Marital status of the clients	{ marr=Married , Nmar=Never married , separ=Separated , divor =Divorced , wido=Widowed , liv2= Living together ,
5	Education	Educational level of the clients	{Illitrate=Illiterate , read=Able to read , prsch= Primary , sesch= Secondary , ugrad=under grad,pgrad=postgraduate
6	Employed	Employment status of the clients	{N=No , Y=Yes }
7	ReaHere	Primary reason the clients visit the center	{Risk , Plan ,status, causal other}
8	PrvTeste	Whether the clients are previously tested	{ No=No ,Yes+ve=Yes HIV positive , “Yes-ve”=Yes HIV Negative , Yes-inc =Yes inconclusive ,NA=Didn’t take
9	EverHSex	Whether the clients have ever had sex	{no=No , yes=Yes}
10	UseCond	Condom usage behavior of clients during sexual intercourse	{Ne=Never , Al=Always, so=Sometimes, Not aplicable=Na }
11	LastSex	whether clients have used condom the last time s/he had sexual intercourse	{N=No , Y=Yes, DN=Doesn’t remember, 99=NA }
12	HistSTI	client’s lifetime history of diagnosed or suspected Sexually Transmitted Infections history	{N=No , yeno=Yes , dnknow=Don’t know , 98=NA}
13	Casual	Number of all those sex partners the client has seen only once or rarely for the past 3 months	{np=No Partner , 1p=One Partner, mp=Multiple Partners}

14	Steady	Number of all those sex partners that the client considers as regular partners for the past 3	{np=No Partner , 1p=One Partner, mp=Multiple Partners}
15	Year	Duration of client in attending lecture successfully in university	{Y1=1_{st} year, Y2= 2_{nd} year ,Y3=3^{rd} year,Y4=4 th year, Y5 fifth year,,Y6= year six , Ny=no year}
16	ExpoTime	Suspected Exposure Time to HIV	{less1m = < 1 month , 1mto3m = 1 to 3 months , 4to6m = 4 to 6 months , greater6m= over 6 months , noexpo =
17	Residence	Clients who came from to the university	Region 1=R1,Region 2=R2…….R14 and No region=other
18	Location	The place where the client is currently living	{Com= commerce, A= not applicable ,other=other, SC=science(Art), SS=social science (sadist kilo),Tec=Technology N .and S.,med= medical(black lion)}
19	Category	Which program the client registerd for learning or not registed student	{Comm=community , EXSTU=extension student,other=other, PART= parent of staff,RESTU= regular student.STAFF=staff, SUSTU= summer student}
20	**HIV**	Laboratory result of HIV status	{neg=Negative , pos=Positive}

Appendix 3: Summary Of Confusion Matrix used for Experimentation

1) J48 algorithms with 10- fold cross validation with all attribute

a b <-- classified as

7040 636 | a = Neg

54 7666 | b = Pos

2) Summary of pruned J48 algorithms with minNumObj = 200 with 10 fold cross validation

a b <-- classified as

6033 1643 | a = Neg

1797 5923 | b = Pos

3) Result of re-evaluation of best J48 decision tree with 3000 dataset

=== Re-evaluation on test set ==User supplied test set Relation: finalt estcaseI nstances: 3000

Attributes: 20

a b <-- classified as

71 8 | a = Neg

71 1388 | b = Pos

4) **The snap shot of Confusion matrix Of PART algorithm with 10-fold cross validation and** all **attribute**

a b <-- classified as

7182 494 | a = Neg

16 7704 | b = Pos

5) The confusion matrix of PART Algorithms with feature selected (9 attribute)

a b <-- classified as

6501 1175 | a = Neg

795 6925 | b = Pos

6) The experimental result of SMO algorithms confusion matrix

a b <-- classified as

1903 437 | a = Neg

458 21 | b = Pos

a) 70-30 percenatge splite

a b <-- classified as

6283 1393 | a = Neg

1593 6127 | b = Pos

b) 10-fold cross validation

Appendix 4. Sample J48 Decision Tree With 10 Fold Cross Validation

J48 pruned tree

```
PrevTest =  yes-ve
|  EverHSex = Yes
|  |  Sex = M
|  |  |  Age = A4
|  |  |  |  MarStatus  = Nmar
|  |  |  |  |  ReaHere = plan
|  |  |  |  |  |  Residance = R14
|  |  |  |  |  |  |  UseCond = Ne: Neg (118.0)
|  |  |  |  |  |  |  UseCond = Na
|  |  |  |  |  |  |  |  Year = Y3: Neg (120.0)
|  |  |  |  |  |  |  |  Year = Y1: Neg (7.0)
|  |  |  |  |  |  |  |  Year = Y4
|  |  |  |  |  |  |  |  |  ExpoTime = dnotkn: Pos (11.0/2.0)
|  |  |  |  |  |  |  |  |  ExpoTime = less1m: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = noexpo: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = 4to6m: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = less1mto3m: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = greater6m: Neg (2.0)
|  |  |  |  |  |  |  |  Year = NY: Neg (17.0)
|  |  |  |  |  |  |  |  Year = Y2: Neg (4.0)
|  |  |  |  |  |  |  |  Year = Y6: Neg (2.0)
|  |  |  |  |  |  |  |  Year = Y5: Neg (1.0)
|  |  |  |  |  |  |  UseCond = So
|  |  |  |  |  |  |  |  HISTI = No
|  |  |  |  |  |  |  |  |  EduStatus = other
|  |  |  |  |  |  |  |  |  |  Category = RESTU: Neg (2.0)
```

Appedix 5: Sample PART Decision Rule List With 10 Fold Cross Validation

PART decision list

PrevTest = yes+ve AND

Steady = 1p AND

Category = RESTU AND

SessType = Indv AND

ReaHere = status AND

ExpoTime = greater6m: Pos (320.0/2.0)

EverHSex = No AND

MarStatus = divor AND

ReaHere = causal: Pos (63.0/1.0)

EverHSex = No AND

MarStatus = divor AND

PrevTest = yes+ve: Pos (46.0)

EverHSex = No AND

MarStatus = wido AND

PrevTest = No: Pos (30.0)

EverHSex = No AND

PrevTest = yes-ve AND

ReaHere = status: Neg (103.0)

PrevTest = yes+ve AND

Category = Comm AND

EverHSex = Yes: Pos (121.0)

PrevTest = yes+ve AND

Steady = 1p AND

Category = 99.0: Pos (291.0/2.0)

EverHSex = No AND

ReaHere = status AND

Year = Y3 AND

ExpoTime = dnotkn: Pos (62.0)

EverHSex = No AND

ReaHere = plan AND

MarStatus = Nmar AND

Residance = R4: Neg (239.0)

EverHSex = No AND

ReaHere = plan AND

MarStatus = Nmar AND

UseCond = Ne AND

ExpoTime = noexpo: Neg (235.0)

EverHSex = No AND

UseCond = Na AND

MarStatus = Nmar AND

EduStatus = Sesch AND

Steady = 1p AND

ReaHere = plan AND

ExpoTime = dnotkn: Neg (169.0)

ReaHere = causal AND

Year = Y3 AND

PrevTest = No AND

EverHSex = Yes AND

Category = RESTU AND

LastSex = Y AND

Age = A4: Pos (75.0)

ReaHere = causal AND

Year = Y3 AND

PrevTest = No AND

EverHSex = Yes AND

UseCond = Na AND

MarStatus = wido: Pos (55.0)

ReaHere = causal AND

Year = Y3 AND

PrevTest = No AND

EverHSex = Yes AND

UseCond = Ne AND

HISTI = No AND

EduStatus = Illitate: Pos (113.0)

Appendix 6: sample selected interesting rule discussed as findings

Rule 1: EverHSex = No AND ReaHere = status AND Year = Y3 AND ExpoTime = dnotkn: Pos (62.0)

Rule 2: ReaHere = causal AND Year = Y3 AND PrevTest = No AND EverHSex = Yes AND Category = RESTU AND LastSex = Y AND Age = A4: Pos (75.0)

Rule 3: ReaHere = causal AND Year = Y3 AND PrevTest = No AND EverHSex = Yes AND UseCond = Na AND MarStatus = wido: Pos (55.0)

Rule 4: ReaHere = causal AND Year = Y3 AND UseCond = Ne AND PrevTest = No AND HISTI = No AND Age = A6: Pos (111.0/3.0)

Rule 5: ReaHere = risk AND LastSex = N AND Year = Y3 AND Residance = R14 AND Sex = F AND EverHSex = Yes AND Age = A6 AND

Rule 6: PrevTest = No AND Age = A6 AND Year = Y3 AND Category = Other AND Sex = M: Pos (106.0/6.0)

Rule 7: PrevTest = yes-ve AND Sex = M AND Age = A3 AND HISTI = NA AND EduStatus = ugrad: Neg (75.0)

Rule 8: Sex = F AND EverHSex = Yes AND Year = Y1 AND UseCond = Ne AND EduStatus = ugrad: Pos (10.0/2.0)

Rule 9: Sex = F AND EverHSex = No AND Residance = R1 AND PrevTest = No: Pos (23.0)

Rule 10: PrevTest = No AND Residance = R14 AND Year = Y3 AND HISTI = No AND UseCond = So AND Age = A3 AND ExpoTime = less1m: Pos (12.0)

Rule 11: Residance = R14 AND HISTI = No AND UseCond = So AND Employed = Yes AND Age = A3 AND ExpoTime = greater6m: Pos (12.0)

Rule 12 :PrevTest = No AND Residance = R14 AND HISTI = No AND ExpoTime = 1mto3m AND Age = A4: Neg (15.0)

Rule 13 : PrevTest = No AND Residance = R14 AND HISTI = No AND LastSex = N AND EduStatus = Sesch AND Employed = NO AND Category = RESTU: Neg (20.0)

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

PREDICTING HIV INFECTION RISK FACTOR USING VOLANTORY COUNSELING AND TESTING DATA: A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)

GIRMA AWEKE

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

PREDICTING HIV INFECTION RISK FACTOR USING VOLANTORY COUNSELING AND TESTING DATA:

A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information science

By

GIRMA AWEKE

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

PREDICTING HIV INFECTION RISK FACTOR USING VOLANTORY COUNSELING AND TESTING DATA: A CASE OF AFRICAN AIDS INITIATIVE INTERNATIONAL (AAII)

By

GIRMA AWEKE

Name and signature of members of the examining Board

Name	**Title**	**Signature**	**Date**
W/ Lemlem Hagos	Chairperson	_____	_____
Dr. Dereje Teferi	Advisor	_____	_____
Dr Rahel Bekele	Examiner	_____	_____

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Girma Aweke

JUNE 2012

The thesis has been submitted for examination with my approval as university Advisor

---

Dr. Dereje Teferi

JUNE 2012

DEDICATION

I would like to dedicate this thesis work to my uncle Abera Dinkayehu who has been struggling and fighting for my education since I was a little boy. Gashe you are always in my heart and your dream is now realized!!

ACKNOWLEDGMENT

First and foremost extraordinary thanks go for my Almighty God and His Mother Saint-Marry.

I would like to express my gratitude and heartfelt thanks to my advisor, Dr. Dereje Teferi for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study.

I am very grateful to the management and staff of AAII, especially Ato Mollalign, for constant support in accessing the data and Ato Zerihun for providing appropriate professional comment, explanation and other supportive document for my research work. My special thanks also goes to Addis Ababa University Library Management for granting me study leave with the necessary benefits, without which I could not have been able to join my M.Sc. study.

I am immensely indebted to my beloved family especially My wife w/o Asnakech Simegn ,my lovely daughter Elishaday Girma, My father Aweke D.,My Mother Almaz N., My Brother Demoz, My sisters Bethlehem, Gete and not mentioned here parents for giving me unconditional care, love, time, patient and support throughout my life.

My special thanks also goes to my friend and classmate Biniam Asnake.Your support and advice has been worth considering starting from the beginning up to the completion of the program.

I also grateful to my friends, staff members of school of Information science(especially Dr. Million Meshesha and other), and my classmate (Tigabu, sewale,Negasi,G/medin,Biazen,Danil M. and other) for encouragement, support and the good friendship they shared.

Last, not least special thanks also goes to Ato Adane Letta and Tesfaye H/mariam(for editing and providing essential comment) , W/t Meseret Ayano for her unlimited support of all relevant information resources whenever her cooperation was needed.

My special thanks also goes to Ato Solomon M , Nebyou A., Abraham G.,Ato Mesfin G., Ato Belay, and Ato Behailu J. who offered good hospitality whenever their cooperation were needed.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAII	African AIDS Initiatives International
AAU	Addis Ababa University
AIDS	Acquired Immunodeficiency Syndrome
ARFF	Attribute-Relation File Format
CDC	Center of Disease Control and Prevention
CHAID	Chi-squared Automatic Interaction Detection
CART	Classification and Regression Trees
CRISP-DM	CRoss-Industry Standard Process -Data Mining
CSV	Comma Separated Values
DM	**Data Mining**
HAPCO	HIV/AIDS Prevention and Control Office
HIV	Human Immunodeficiency Virus
KDD	Knowledge Discovery in Database
MoH	Ministry of Health
UNAIDS	United Nations AIDS
UNFPA	United Nations Population Fund
VCT	Voluntary Counselling and Testing
WHO	World Health Organization
WEKA	Waikato Environment for Knowledge Analysis

Table of Contents

ABSTRACT

Despite a great deal of efforts, the world still has neither a cure nor a vaccine for HIV/AIDS infection. Millions of people have been suffering from this incurable disease. Fortunately, researchers have become successful in prolonging and improving the quality of life of those infected with HIV. Nonetheless, it has become increasingly clear that preventing the transmission and the acquisition of HIV through educating people to bring about behavioral changes should be the focus.

The widely and freely available voluntary counseling and testing center (VCT) in Addis Ababa which provides an enormous role in counseling, promoting and checking clients HIV status through the clinical laboratory test. In line with this, the center has been collecting client's information or records for further investigation with confidentiality. The record consists of many attribute that may have a direct or indirect impact with HIV infection. Moreover, identifying HIV infection risk factors or determinate variables provides benefits at different level of the society (such as individual, community and organizational level).

The benefit not yet known by the client rather the organization keeps their records after they got tested. To this end, great efforts have been made to develop models to identify HIV infection risk factor using data mining technology.

This research is initiated to identify the determinant risk factors of HIV infection by developing predictive models to support voluntary counselling and testing service of African AIDS Initiatives international (AAII) provided at Addis Ababa University and its surrounding.

The six steps hybrid methodology has been followed for predictive HIV infection risk factors modeling among selected attributes. Three classifications techniques such as Decision tree J48, PART and SMO algorithms were experimented for building and evaluating the models.

Before experimentation data pre-processing task has been performed to remove outliers, fill in missing values, and select best attributes, discretization and transformation of data. The pre-processing phases took considerable time of this work. A total of 15,396 VCT client records have been used for training the models, while a separate 3,000 records were used for testing their performance. The model developed using the PART algorithm has shown the best classification accuracy of 96.7%. The model has been evaluated on the testing dataset and scores a prediction accuracy of 95.8%. The results of this study have shown that the data mining techniques were valuable for predicting HIV infection risk factors. Hence, future research directions are forwarded to come up applicable solutions in the area of the study.

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Despite a great deal of efforts, the world still has neither a cure nor a vaccine to HIV/AIDS (Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome) infection. Millions of people are suffering from this incurable disease. Fortunately, researchers have become successful in prolonging and improving the quality of life of those infected with HIV. Nonetheless, it has become increasingly clear that preventing the transmission and the acquisition of HIV through educating people to bring about behavioral changes should be the focus (UNAIDS, 2010).

Although the United Nations Aids Department (UNAID) has reported that the HIV prevalence is leveling off and there is a fall in the numbers of new infections globally (UNAIDS, 2010), it remains one of the leading obstacles to health and growth of developing countries. There are still a huge number of people infected and affected by HIV. For instance, in 2009, there were an estimated 2.6 million people who became newly infected with HIV. The incidence of HIV infection declined by 19% between 1999 and 2009 globally; the decline exceeded 25% in 33 countries, including 22 countries in sub-Saharan Africa. However, it was also reported that HIV/AIDS has become the leading cause of death in the world (UNAIDS, 2010).

Ethiopia is one of the Sub-Saharan African countries most severely affected by the HIV/AIDS pandemic. Currently, the national adult prevalence rate is estimated at 2.3 percent and an estimated number of 1.2 million people are living with HIV/AIDS , an estimated 67,000 people lost their lives due to AIDS at the end of 2007 (UNAIDS/WHO, 2010). The Federal HIV/AIDS prevention and control office also report that young women were particularly vulnerable to HIV infection as compared to young men. Among women aged 15-19 and 20-24, for example, 0.7 and 1.7 percent, respectively, were HIV-infected, while the figure for men in the same age categories showed that only 0.1 and 0.4, respectively, were infected (FDRE HIV/AIDS PC, 2010).

Despite the overwhelming challenges, the Government of Ethiopia is making efforts towards containing the epidemic and mitigating the impact of HIV/AIDS through an intensified national response in a comprehensive and accelerated manner.

To this end, huge effort had been made to build the implementation capacity, especially in the health sector, in areas of human resource development, site expansion and construction of health facilities. There are various non- profit organizations that have the mission of preventing HIV infection, improve care and support and build capacity to address the global HIV/AIDS pandemic. Among them, African AIDS Initiative International (AAII) is the leading international organization founded on April 2001 officially by opening a regional office in Addis Ababa, Ethiopia (LLGP, 2007).

The office plays an enormous role in preventing the spread of HIV/ AIDS in the university community, in this regard, the statistics about Voluntary Counseled and Tested (VCT) service of AAU indicates that the number of treatment seekers and counseling test services has highly increased and currently estimated 18,800 clients have taken the service. Based on a recent research report of Nigatu and Seman (2011) out of the total respondents, 47.2% had been tested for HIV/AIDS and more than 80% had ingress to take VCT service for HIV/AIDS.

Additionally, the attitudes and practices of the Addis Ababa University community towards HIV prevalence has been exposed by the risk factors such as sex, previous place of residence, religious participation, pornographic film show, alcohol intake, chat chewing, and Cigarette smoking. A comprehensive approach to prevent HIV includes, creating an opportunity for people to participate in VCT service (Nigatu and Seman , 2011). Besides counseling and testing service, it is best to investigate the degree of association between the risk factor mentioned above which is equally important in eradicating HIV infection.

1.1.1. Data Mining and Health care

Widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer-assisted analysis (Larvac, 1998). Extensive amounts of data gathered in health care databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. It has been estimated that an acute care hospital may generate five terabytes of data a year (Fayyad et al., 1996). The ability to use these data to extract useful information for quality health care is crucial.

Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified; efficiency and accuracy of decisions decrease when humans are put into stress and immense work (Eapen , 2004). Prather et al. (2001) wrote that there are only a few tools to evaluate and analyze clinical data after it has been captured and stored. The researcher further stated that evaluation of stored clinical data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management.

The traditional method of turning data into knowledge relies on manual analysis and interpretation (Fayyad et al., 1996). Prather et al. (2001) argued that to evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data, to discover new patterns and relationships hidden in the data. It is due to these challenges of searching for knowledge of relational databases and our inability to interpret and digest these data as readily as they are accumulated, which has created a need for a new generation of tools and techniques for automated and intelligent database analysis. Consequently, the discipline of knowledge discovery in databases (KDD), which deals with the study of such tools and techniques, has evolved into an important active area of research (Raghavan et al., 2002).

Data mining is one among the most important steps in the knowledge discovery process. It can be considered the heart of the KDD process. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the data (Eapen, 2004). It is a new generation of computerized methods for extracting previously unknown, valid, and actionable information from large volume of database and then using this information to make critical decision (Cabena et al., 1998). Hence, it is better to explore the relevance of this technology to any organization that performs health-related activity.

There are various techniques and tools of data mining that help researcher in exploring the benefit of data mining in various sectors. Data mining techniques can be broadly classified based on what they can do, namely description and visualization; association and clustering; and classification and estimation, which is predictive modeling. Predictive modeling can be used to identify patterns, which can then be used to predict the odds of a particular outcome based upon the observed data.

Rule induction is the process of extracting useful if/then rules from data based on statistical significance. A decision tree is a tree-shaped structure that visually describes a set of rules that caused a decision to be made. For example, it can help determine the factors that affect kidney transplant survival rates and the nearest-neighbor method (DMSHO, 2011). Each of these techniques analyzes data in different ways.

In addition to data mining technique mentioned above, most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools (Silltow, 2006). According to Kon and Geraled (2005) there is vast potential for data mining applications in healthcare. Generally, these can be grouped as the evaluation of treatment effectiveness; management of health care; customer relationship management; detection of fraud and abuse; and the identification of risk factors associated with the disease.

1.2. Statement of the Problem

The costs of fatalities and severity due to HIV/AIDS have a tremendous impact on societal well-being and socioeconomic development. HIV/ADIS is among the leading causes of death worldwide, causing an estimated 1.8 million deaths per year (UNAIDS, 2010). According to the report of the second round HIV/AIDS Behavioral Surveillance Survey in Ethiopia, it was found out that around 9.9 percent of the in-school youth (14.6 % of males and 5.3 % of females) had sexual experience (BSS, 2005). Nigatu and Seman (2011) in their research reported that HIV/AIDS is affecting young members of the societies especially adolescents between the age of 15 to 24 and proved that these age groups were vulnerable and at risk of the disease.

A number of studies have also shown that AIDS has progressively been on the increase and constitutes a big problem among college and university students, although the extent of the problem is relatively unknown (Abdinasir and Sentayehu, 2002; Elias, 2009; Tefera, Challi and Yoseph, 2004.; Teka, 1993; Getinet, 2009; AAII, 2006). Evidence showed that most sexual risk behaviors among college and university students might have been acquired through a period of campus life (Teka, 1993). This may be due to the life of independence, away from parental control, that often characterizes such a setting.

University students are often viewed as being at high risk for HIV infection due to their propensity to engage in exploratory behavior and their needs for peer social approval and false sense of non-vulnerability (Beyene, Solomon and Yared, 1997).

A number of factors contribute to the cause of HIV infection such as economic, social, behavioral, environmental, educational and health related factors (Nigatu and Seman , 2011). Among these there are variable or attribute that have trivial and non-trivial role in causing a client being infected with HIV. Hence, categorizing and/or predicting HIV infection risk factor of attribute or variable that have causal relation or pattern from voluntary counseling and testing data is very crucial for organizations who are investing large amounts of money on preventing and controlling HIV/AIDS.

In attaining its objective, AAII collects daily, monthly, quarterly and yearly health-related statistics of the university community (LLGP, 2007). It analyzes the available data with traditional statistical tools or technique alone is not enough for health professional, planner and policy maker to identify major determinant risk factors for HIV infection and transmission. As Plate et al. (1997) indicated that traditional method of data analysis has limited capacity to discover new and unanticipated patterns and relationship that are hidden in conventional databases. Identifying patterns of attribute or risk factors of HIV infection is difficult because the dataset contains or involves too many attribute or parameter. So studying the patterns of the attribute that have relationship among each other rather than listing are more crucial.

To the knowledge of the researcher, no attempt has been made using rule induction and support vector machine classification technique to identify the determinant factor of HIV infection. Moreover, this work has been the first attempt to evaluate the applicability of DM for predicting HIV infection risk factor in higher education institutions. Last not least, three unique variables (Category, campus and Year of the student) have been used as an additional attributes for analysis of the experimentation that was not used in the previous studies.

Therefore, the purpose of this study is to investigate the underlying determinant risk factors of HIV infection from the available VCT data using data mining techniques. To this end, it has been attempted to obtain answer for the following research questions:

- What are the main determinate risk factors (attributes) of Voluntary counseled and tested client records that cause HIV infection in the Addis Ababa University community and their surroundings?
- What are the most interesting patterns or rules generated/predicted to determinate risk factor for HIV infection that can be used as a cause of HIV /AIDS.
- To what degree (accuracy level) the determinate risk factor for HIV infection be determined by applying data mining.

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this study is to apply data mining technology to identify and predict the major risk factor of HIV infection using VCT records by developing a model that can support counseling and testing service provider, policy makers, and planners.

1.3.2. Specific Objectives

For the realization of the general objective stated above, the following specific objectives have been formulated.

- To review the literature on DM technology and their application in the Health care Industry particularly in HIVAIDS for the purpose of getting information that will help in this research
- To prepare the data for analysis and model building, by cleaning and transforming the data into a format suitable for the selected DM algorithms
- To identify the features of risk factor for HIV infection, and select the appropriate classification algorithms to be used based on the type of data and objectives of the study
- To train and develop a classification model that support in predicting HIV infection risk factor
- To test and compare the resulting performances of the classification models and recommend the overall best results of the classification models.
- To report findings of the result and forward recommendation for further research.

1.4. Research Methodology

According to Smolander et al. (1990) a method can be considered as a predefined and organized collection of techniques and a set of rules which state by whom, in what order, and in what way the techniques are used to achieve or maintain some objectives. The DM process model describes procedures that are performed in each of its steps (Jinhong et al. 2009). It is primarily used to plan, work through, and reduce the cost of any given project.

1.4.1. Research Design

This research is designed to identify the determinate risk factor of HIV infection. To explore the application of data mining on this particular research, hybrid (Ciso.et al) data mining methodology was employed. This model was developed, by adopting the CRISP-DM model to the needs of academic research community (Fayyad, 1996). Unlike the CRISP-DM process model, which is fully industrial, the hybrid process model is both academic and industrial. Additionally, it extends it functionality in to research oriented description of the steps; incorporating DM step instead of the modeling step, and increasing number iteration among steps to make the research process more complete or to come up with explicit feedback mechanisms.

Ciso.et al involves sixth iterative process or steps including: understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge steps (Pete, 2000).

As depicted in figure 1.1 to determine HIV infection risk factor form VCT dataset data mining tools is not the only tools that we require instead review of literature(conceptual and theoretical) be done. Moreover, we need to understand the system that contain the initial data and other intermediate data processing tools and technique Such as MS Excel and Epi Info software to transfer and Pre-process the data for actual application of data mining task using Hybrid data mining model. The researcher tries to discuss each step in details below.

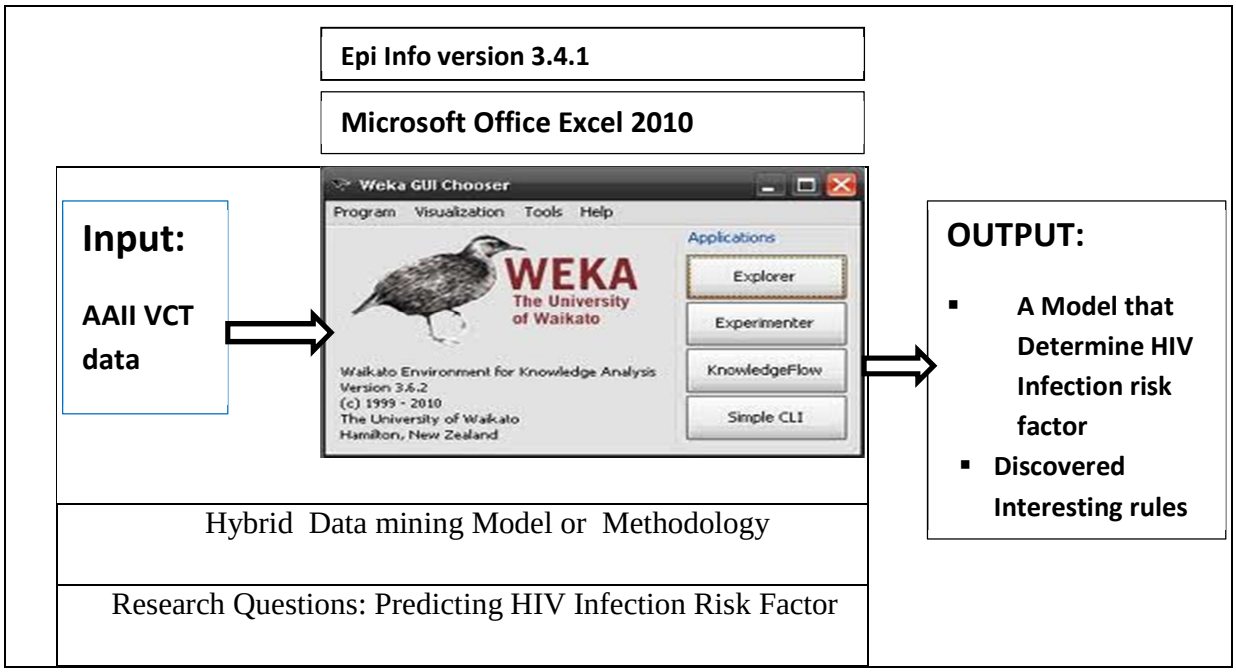


Figure 1. 1 Diagrammatic Skeleton of overall Research Methodology and Design

1.4.1.1. Understanding of the problem domain

The primary task of the researcher in knowledge discovery process is a close look of the problem domain. To identify, define, understand and formulate the problem domain (VCT service problems), various data collection methods were used including: interview, observation, document analysis. In-depth interview has been conducted with domain expert to determine attribute feature selection and understand some complex business process. As a result of insight gained knowledge of the domain data mining problem is defined. Further, databases such as VCT data, major service AAII provides and HIV/AIDS related policy and related issue are consulted to gather the pertinent data for the present research.

The main purpose of the AAII VCT centers is to save the life of the university community and surrounding particularly the industrious people by mitigating the prevalence of HIV/AIDS through establishing a permanent VCT center in each campus and surroundings.

1.4.1.2. Understanding the VCT data

Once the problem to be addressed is understood, the next step is analyzing and understanding the data itself. Because, the end result of knowledge discovery process heavily depends on the quality and quantity of available data (Cios et al., 2007). In this phase, the original data collected for this research should be described briefly. Its description includes listing out attributes with their respective values, missing and outlier and evaluation of their importance to the research goal.

Additionally, careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the DM tasks to be performed.

1.4.1.3. Preparation of the data

This is the most crucial phases in which the success of the entire knowledge discovery process depends. It usually consumes much of the entire research effort. In this phase, the final dataset is also constructed from the initial raw data. Data preparation tasks were performed in iterative ways. The major tasks include: description of data sources, carrying out statistical summary measure, finding out distinct value, filling missing values, outlier and noisy data and data transformation/reduction activities were also undertaken in this phase. Additionally, feature / attribute construction was also made too. The researcher decides, together with domain experts, the data with respective attribute that is used as input for applying the DM techniques (see Appendix 2).

Data cleaning (or data cleansing) routines was applied to fill in missing values (with the most frequented or modal value), smooth out noise (by removing the record), and detect outliers (by removing or substituting with moral values) in the data. To support the preprocessing tasks, the researcher used WEKA 3.7.5 to normalize and to fill missing value with modal value. It also helps to train the selected algorithms with training sample and testing case.

1.4.1.4. Data mining

The prerequisite for the application of data mining research goal is preprocessing task. Besides, better understanding and preparation of data result in a good selection of data mining methods

otherwise it leads the miner to incorrect decision (Fayyad, 1996). So this step is all about selecting appropriate data mining methods based on the research objectives.

The use of DM to gain knowledge discovery regularities in the data and not just predictions has been common (Witten and Frank, 2000). We can understand from the objectives of the research, gaining knowledge and discovering patterns that cause HIV infection within VCT data is certainly the purpose of the study. Consequently, a classification technique particularly decision tree, rule induction and support vector machine techniques were selected and used for prediction of the HIV infection risk factor. The researcher selected only classification technique because the research dataset has clear and simplified labeled class. For training and implementation purpose J48, PART and SMO algorithms were used respectively.

1.4.1.5. Analysis and Evaluation of the discovered knowledge

In data mining evaluation has two primary functions. Primarily, it helps to predict how well the final model will work in the future. Secondly, evaluation is an integral part of many learning methods and helps to explore the model that best represents the training data. According to Ciso et al. (2007) the novelty and interestingness of the models are evaluated with the agreement of the researcher and domain expert.

The different classification models developed in this research were evaluated using a test dataset based on their classification accuracy. In the other word evaluation of the performance of the classifier is also made in terms of different confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), Relative Operating Characteristics (ROC)), the number of correctly classified instances, number of leaves and the size of the trees, execution time.

Furthermore, models can be compared with respect to their speed, robustness, scalability, and interpretability which may have an influence on the model (Kantardzic, 2003). The confusion matrix of the classifier model was analyzed in terms of the following variables in Table 1.1

Table 1. 1 Sample summary of Confusion Matrix of the model

		Predicted the status of HIV infection		Total
		Negative	Positive	
Actual HIV Status	Negative	TN	FP	TN + FP
	Positive	FN	TP	TP + FN
Total		TN + FN	FP+ TP	TN + FP + FN + TP

Key: TN = True Negative FP= False Positive FN= False Positive TP= True Positive

Contextually what each variable stands for is stated as follows, TN: The number of HIV negative clients that are classified as negative. FN: The number of HIV positive clients that are classified as negative. TP: The number of HIV positive clients that are classified as positive. FP: The number of HIV negative clients that are classified as positive. FP + TN: The total number of actual HIV negative clients. TP + FN: The total number of actually HIV positive clients. TN + FN: The total number of predicted HIV negative clients. FP+ TP: The total number of predicted HIV positive clients. TN + FP + FN + TP: The total number of all clients.

The performance of the experiments is measured in the following manner as depicted below:

Sensitivity (TPR): indicated by the proportion of HIV positive clients that are correctly classified as positive i.e. TPR= TP/ TP + FN

Specificity (TNR): The proportion of HIV negative clients that are correctly classified as negative. TNR= TN/ TN+FN

False Negative Rate (FNR): The proportion of HIV positive clients that are erroneously classified as negative i.e. FNR= 1 – TPR or FNR = FN/ TP + FN hence, the sum of TPR plus FPR equals 1.

Correctly Classified Instances (Accuracy): To compute the proportion of clients those are correctly classified using this formula i.e. **Accuracy** = (TP + TN) / (TN + FP + FN + TP)
Similarly, Incorrectly Classified Instances (Error Rate): The proportion of clients that are incorrectly classified. Error Rate = (FP + FN) / (TN + FP + FN + TP).

Furthermore we can also compute the effectiveness and efficiency of the model in terms of recall and precision. As a result, recall can be computed for both positive and negative classes. So the formula of the recall for negative class be TN/TN + FP and for positive class TP/ TP+FN. Similarly to compute the perception of the model the following formula be used for negative and positive class respectively, TN/ TN + FN and TP/ FP+ TP.

Besides, this research plans to use the following methodology to develop data mining model that determines the risk factor for HIV infection.

1.4.2. Literature review

An in depth literature review is conducted by refereeing of books, journal articles, conference paper and the internet get more insight to the concept of data mining and its application, especially in health related areas.

1.4.3. Data identification, collection and preparation

The primary source of data to conduct this research was voluntary counseled and tested database of AAII which contains more than 18,800 records. To this end, the data are collected and arranged into new database to make it suitable for the experiment and for the selected data mining tools. Therefore, a new database was prepared by analyzing the collected data using preprocessing tasks like data cleaning and reduction.

1.4.4. Model building

As indicated in the research objective the researcher used data mining technique to develop a model that identify determinant and predict risk factors for HIV infection. So, in the experimental part predictive models has been developed by using the selected data mining technique. The

research was conducted with three selected techniques of data mining. Finally, a comparison of them was made to get a reasonable accuracy.

1.4.5. Tools and Software

The following tools and application software have been used to accomplish the research process: **WEKA 3.7.5 data mining tools**: Based on the prior knowledge of the researcher on the applicability of the tools for the purpose and freely availability of the software made the researcher to use the tools to build analysis and evaluate the model being developed for research goal.

Ms-Excel: it was used for data preparation, pre-processing and analysis task because it has the capability of filtering attribute with different values. Besides, it is a very important application software to make ready the data and easily convert into of the file.

Ms-word and Ms Power point: It is a known fact that the final report document is prepared with word and for presentation purpose preparing slide is inevitable. So the use of power point is mandatory.

Epi info: this is standard software that used to handle health related data. As a result, AAII handles its VCT data with it. So to pre-process the original data found from it the researcher employed this software specifically to calculate missing value and detect outlier.

1.5. Scope and Limitation of the Study

The main aim of this research is to find out the applicability of DM for determining and predicting HIV infection risk factor in the VCT provider center in Addis Ababa. This research focuses only on client records of Addis Ababa University (AAU) community and its surrounding (Arada sub-city) for the reason that university community is more exposed to those places to HIV/AIDS. Specifically, 15,396 records of data for knowledge discovery are obtained from the African AIDS Initiatives International (AAII).

The dataset of AAII also contains other higher education institution's client records. However, it is not included in this research. Among the 63 attributes, 20 of them are used in experimentation, which are selected by the domain experts.

In this study, we evaluated three classification techniques to develop a prediction model to identify major determinant risk factors for HIV infection.

1.6. Significance of the Study

Though the primary goal and initiatives of this research has been for academic exercise, the findings of this research can ultimately use in various areas.

The organization (both AAII and AAU) may use the predictive model to determine risk factor of HIV infection based on the available VCT data and understand the potential of each variable towards causing HIV infection in higher educational institutes particularly in Addis Ababa University. This helps to design and develop strategies for the efforts towards struggling against HIV/AIDS care, control and prevention program. To this end, the organization protects and save the lives of thousands of peoples.

The government, Ministry of Health, Ministry of education, the university management and policy maker can also use the output of the research as a decision support for formulating policy related to HIV care, control and prevention Program in higher education institutes.

The research also gave opportunity for utilization of health care data particularly VCT test data that is stored for no purpose , and the model developed also enables less qualified data manger or health professional staff assigned to perform the task of determining and managing HIV infection more accurately and confidently and less subjectively.

Last not least, the output of the research also contribute to understanding of the theoretical views and practical problems in the design and implementation of VCT programs and invite interested researchers to explore more in related and similar areas.

1.7. Organization of Thesis

The research report in this study has six chapters. The first chapter deals with the basic overview including background, statement of the problem, objective, scope of the study, methodology and thesis organization.

In the second chapter critical literature review which includes review of national and global research and trend of HIV/AIDS in higher education. Definition , benefit of VCT services has been discussed.

In the third chapters review of data mining literatures was presented. Definition, evolution, types, application of data mining in different domain areas was discussed. Review of related work particularly application of data mining in healthcare sector especially with the issue of HIV/AIDS was discussed here.

The fourth Chapter deals with data preprocessing tasks. In this chapter how the major data preprocessing tasks were applied to the current data were shown. Data cleaning, reduction and preparation of dataset to be used as input for predictive model.

The fifth chapter deals with experimentations and result interpretations. In this chapter building of model with training dataset and validating the result with testing datasets ,and interpretation of the result of the experimentation were the major concern. Finally, a comparison of the algorithms used for reasonable accuracy was made.

In the sixth chapter conclusions and recommendations were presented

CHAPTER TWO

HIV/AIDS AND CURRENT RESEARCH OUTPUT

Under this chapter attempt has been made to present a review of the literature and trends of HIV/AIDS globally as well as locally, overview, process and benefit review of voluntary counseling and testing (VCT) services in Ethiopia is also made. Lastly, application of data mining in healthcare particularly in HIV/AIDS is thoroughly reviewed. These help a great deal in understanding the domain area where data mining technology is going to be used.

2.1. Overview of HIV/AIDS

HIV, the virus that causes AIDS (Acquired Immunodeficiency Syndrome) (UNAIDSb., 2010) has become one of the world's most serious healthiest and development challenges. The first cases were reported in 1981 (AIDS, n.d.). Since then, the disease has spread rapidly across the globe; in 2010 sub-Saharan Africa was the most HIV infected region in the world with nearly 22.9 million (67%) people living with HIV in the world which is nearly two-thirds of the global burden [UNAIDSa ,2011).In Africa as a whole, the death rate in 2010 is reported to be 1.2 million [UNAIDS ,2011).

These figures are more staggering when you look at the results alongside findings that only 12% of men and 10% of women in the general population had been tested for HIV and received their results (WHOa, 2010). In addition, many of the men and women who seek HIV testing and counseling, are already in the advanced stages of the disease [Hogg, et al. 2006]. The WHO and UNAIDS proclaim that "early diagnosis presents an opportunity to provide people with HIV the information and tools to prevent HIV transmission to others" (WHOb,2004).

A voluntary counseling and testing (VCT) program has been formulated by the WHO and UNAIDS to encourage people to be pre-HIV test counseled, tested and post-HIV test counseled in an endeavor to prevent infection and transmission of HIV.

2.2. Global Issue of HIV/AIDS Epidemic

Global HIV/AIDS data show that the HIV/AIDS epidemic is a worldwide public health problem of great magnitude. About two decades since the beginning of the HIV/AIDS epidemic, over 34 million people worldwide have currently living with HIV and nearly 30 million people have died of AIDS-related causes since the beginning of the epidemic (UNAIDSa, 2011).

Although cases have been reported in all regions of the world, almost all those living with HIV (97%) reside in low- and middle-income countries, particularly in sub-Saharan Africa(UNAIDSa ,2011). Most people living with HIV or at risk for HIV do not have access to prevention, care, and treatment, and there is still no cure (WHOa,2010)

HIV primarily affects those in their most productive years; about half of new infections are among those under age 25 (UNAIDSa,2011). HIV not only affects the health of individuals, it impacts households, communities, and the development and economic growth of nations. Many of the countries hardest hit by HIV also suffer from other infectious diseases, food insecurity, and other serious problems. An evidence of this that people living with HIV at the end of 2010, up from 28.6 in 2001 to 34 million (UNAIDSc, 2011).

Despite these challenges, new global efforts have been mounted to address the epidemic, particularly in the last decade, and there are signs that the epidemic may be changing course. According to the Global HIV/AIDS Epidemic report (UNAIDSa, 2011) depicted in Table 2.1, even though, people living with HIV increase globally, deaths have declined due to the scale up of antiretroviral treatment (ART).

Besides that, Table 2.1 indicates that the global prevalence rate (the percent of people whose ages ranging from (15–40) are infected) was 0.8% in 2010 (UNAIDSa ,2011). According to the report of UNAIDS(2010) 1.8 million People died of AIDS , which indicate a 21% decrease since 2005, and new HIV infections rate have declined by more than 20% since their peak in 1997, and declined by 15% between 2001 and 2010. Still, there were about 2.7 million new infections in 2010 or more than 7,000 new HIV infections per day (UNAIDS, 2011).

2.2.1. Global Response of HIV Epidemic

With this regard, fight towards HIV/AIDS pandemic must adopt an approach that emphasizes the collective responsibility of individuals, community groups, different levels of government and other agencies to diminish irredeemable diseases (GHIVR, 2011).

Region	Total No. (%) Living with HIV end of 2010	Newly Infected in 2010	Adult Prevalence Rate 2010
Global Total	34 million (100%)	2.7 million	0.80%
Sub-Saharan Africa	22.9 million (67%)	1.9 million	5.00%
South/South-East Asia	4.0 million (12%)	270,000	0.30%
Eastern Europe/Central Asia	1.5 million (4%)	160,000	0.90%
Latin America	1.5 million (4%)	100,000	0.40%
North America	1.3 million (4%)	58,000	0.60%
Western/Central Europe	840,000 (2%)	30,000	0.20%
East Asia	790,000 (2%)	88,000	0.10%
Middle East/North Africa	470,000 (1%)	59,000	0.20%
Caribbean	200,000 (0.6%)	12,000	0.90%
Oceania	54,000 (0.2%)	3,300	0.30%

Table 2. 1 HIV Prevalence & Incidence by Region (extracted from UNAIDSa ,2011)

As a result, from the beginning of the 21st century, the international community identified and declared HIV as the formidable health and development challenge. A rapidly expanding HIV epidemic was already dramatically reversing decades of progress on key development indicators, such as infant mortality and life expectancy (UNAIDSb, 2010).

Although the global incidence of HIV infection had peaked in the mid-1990s, more than 3 million people were being newly infected per year, AIDS had become one of the leading causes of adults dying in sub-Saharan Africa and the full assault of the epidemic would not be felt until 2006, when more than 2.2 million people died each year from AIDS-related causes (UNAIDSP,2011).

The revolution in HIV treatment brought about by combination of antiretroviral therapy in 1996 had forever altered the course of disease among those living with HIV in high-income countries but had only reached a fraction of people in low and middle-income countries, which bore 90% of the global HIV burden (UNAIDSb, 2010).

According to the report of 13th International AIDS Conference in July 2000 in Durban, South Africa (see details at www.nlm.nih.gov/news/aidsnote00.html), activists, community leaders, scientists and health care providers joined forces to demand access to treatment and an end to the enormous health inequities between the global North and global South. Months later, world leaders established the Millennium Development Goals, a series of ambitious, time-bound targets aimed at achieving progress on several health and development goals over the next 15 years, including Millennium Development Goal: combat HIV, malaria and other diseases (GHSSH, 2011).

In 2001, the United Nation General Assembly Special Session on HIV/AIDS approved commitment with common targets in specific technical areas, such as expanding access to antiretroviral therapy, antiretroviral prophylaxis to prevent the mother to- child-transmission of HIV and HIV prevention (Schneider et al., 2010).

The member of the general assembly declared a dedicated global health fund to finance the HIV response, resulting in the launch of the global fund to fight AIDS, tuberculosis and malaria one year later: The global fund quickly became a cornerstone in the global response to HIV, funding country-led responses through a pioneering, performance-based grant system.

According to Schneider et al. (2010) in 2003 the United States Government announced the United States President's Emergency Plan for AIDS relief. At US$ 15 billion over five years, it was the largest single funding commitment for a disease in history. The United States President's Emergency Plan for AIDS Relief was reauthorized in 2008 for up to US$ 48 billion to combat AIDS, TB and malaria for 2009–2013 and supports strategic interventions in the drugs and diagnostics markets in 94 countries. Besides, the global HIV/AIDS response indicates (GHIVR, 2011) that there is an immediate response politically as well as financially towards against HIV epidemic. The commitment is in parallel with normative guidance and strategic technical

innovations, including a ground-breaking approach to scaling up treatment access in low- and middle-income countries: the public health approach to antiretroviral therapy.

To mention the vital elements of the public health approach includes: using standardized treatment protocols and drug regimens; simplified clinical monitoring; maximizing coverage with limited resources; optimizing human resources for health and involving people living with and affected by HIV in designing and rolling out antiretroviral therapy programs (USAIDHPI, 2010).

As mentioned in the general fact of HIV/AIDS across the world, it is not the right time to keep silent about this incurable disease. However, everybody should do or react something against HIV/AIDS. For same reason, the researcher of this experimental study is striving from part towards the fight against HIV by implementing the possible and available Technology to the problem domain.

2.3. The Status of HIV/AIDS Epidemic in Ethiopia

In Ethiopia HIV/AIDS infections were first identified in 1984, and the first AIDS cases were reported in 1986 (Garbus, 2003). The prevalence of the disease was low in 1980s, but it escalated quickly through the 1990s. As a result, it rose from an estimated 3.2% of the adult population in 1993 to 7.3 % by the end of 1999(MOH, 2000).

According to NIC (2002) Ethiopia is classified (along with Nigeria, China, India and Russia) as belonging to the 'next wave countries' with large populations at risk from HIV infection which eclipse the current focal point of the epidemic in central and southern Africa. It is estimated that seven to 10 million Ethiopians be infected by 2005 because of the current high adult prevalence rate, widespread poverty and low educational levels (Garbus, 2003).

The dominant mode of transmission is through heterosexual contact (estimated to account for 87% of infections) and mother to child transmission (MTCT) (10% infections) (GoE, 2004). However, the prevalence of HIV/AIDS has reduced gradually. Ethiopian Ministry of Health estimates that the current adult HIV prevalence decline to 3.5% (MOH,2006). This figure jumps to an estimated 5% among pregnant women, but uptake of antiretroviral prophylaxis for Preventing Mother-to-child Transmission of HIV (PMTCT) has been minimal and the rate of HIV transmission to

children born to HIV positive women remains at 25 percent. According to the report of Ministry of health two million Ethiopian peoples are living with HIV-positive, but among this it is estimated that fewer than 10% know their HIV status (MOH, 2006).

The above report suggests that the percentage of prevalence rate increase in 1990s from three to seven, but Since 2005 HIV prevalence rate was reduced to 3.5 % (MOH, 2006). According to the current report of UNAIDS, Ethiopia has an estimated two million people living with HIV and the third highest number of infections in Africa. With a population of 83 million people and per capita income of less than US$100 annually, it is also one of the world's poorest countries (UNAIDSE, 2011).

2.3.1. Government Response Towards HIV/AIDS Epidemic in Ethiopia

The government's response to HIV/AIDS was immediate, but inadequate. A National Task Force was established in 1985 following the report of HIV prevalence in the country. Efforts were made, although on a limited scale, to expand information, education and communication (IEC), condom promotion, surveillance, patient care, and HIV screening laboratories at different health institutions(Hailom et al.,2005).

As part of these efforts to address the epidemic, the Government of Ethiopia has expanded HIV/AIDS Prevention and Control Offices at national, regional, zonal and sub-city/wereda levels. The overall objectives of the National office are to guide the implementation of the successful programs to prevent the spread of HIV/AIDS, decrease vulnerability of individuals and communities to the epidemic, provide care services for those living with the virus and reduce the adverse socio-economic consequences of the epidemic (NAC, 2001).

Besides that, the Ethiopian Ministry of Health has set a strategic plan for year 2007 targets, which include: encourage individuals to receive HIV counseling and testing; Address and create opportunity for those people who has getting HIV positive to be on antiretroviral treatment and give particular attention for HIV positive pregnant women to be treated with a complete course of antiretroviral prophylaxis to prevent transmission from mother to child (MOHHA, 2006).To achieve this goal, the Ministry of Health has an ambitious roadmap for scale-up of HIV prevention,

treatment and care that plans the delegation of responsibility for achievement of national targets from central hospitals outwards to the local health centers and private clinics.

Although there is a challenge to address evenly HIV prevention, testing and care in Ethiopia. Eighty-five percent of the population lives in rural areas and suffers from a severe lack of access to public health services. There is also a critical shortage of physicians (an estimated 1,200 in public service practice for a population of 83 million) and other trained health care workers. In addition to this, the per capita expenditures for health from all sources is only US$5.60 compared to US$12.00 per person in the Africa region as a whole (Hailom et al.,2005).

In response to the extreme shortage of qualified health care workers, the Ministry also enacted a plan to further decentralize its operations through collaborations with community-based NGOs. The Ministry of Health also encouraged African Services to replicate its highly successful prevention outreach and VCT model at additional sites and to add diagnosis and prophylaxis of opportunistic infections and antiretroviral therapy to its VCT services.

2.4. Voluntary HIV Counseling and Testing Service

A greater knowledge of HIV status within a community is critical to expanding access to HIV treatment, care and support in a timely manner as it offers people living with HIV an opportunity to receive information and tools to prevent HIV transmission to others [WHOa 2010].

The most successful method of gaining information about HIV has been through Voluntary Counselling and Testing (VCT). According to Mugula et al.(1995) Voluntary HIV counselling and testing is the process by which an individual undergoes counselling enabling him or her to make an informed choice about being tested for HIV. This decision must be entirely the choice of the individual and he or she must be assured that the process be confidential. The scope of VCT service includes :HIV pre-test counselling, testing, and post-test counseling. It is an entry point for prevention and care (FHI, 2004).

2.4.1. The process of VCT service

In HIV counseling, although the goal is eliminating risk, it is recognized that this can be best achieved through small steps for incremental behavior change that bring about risk reduction.

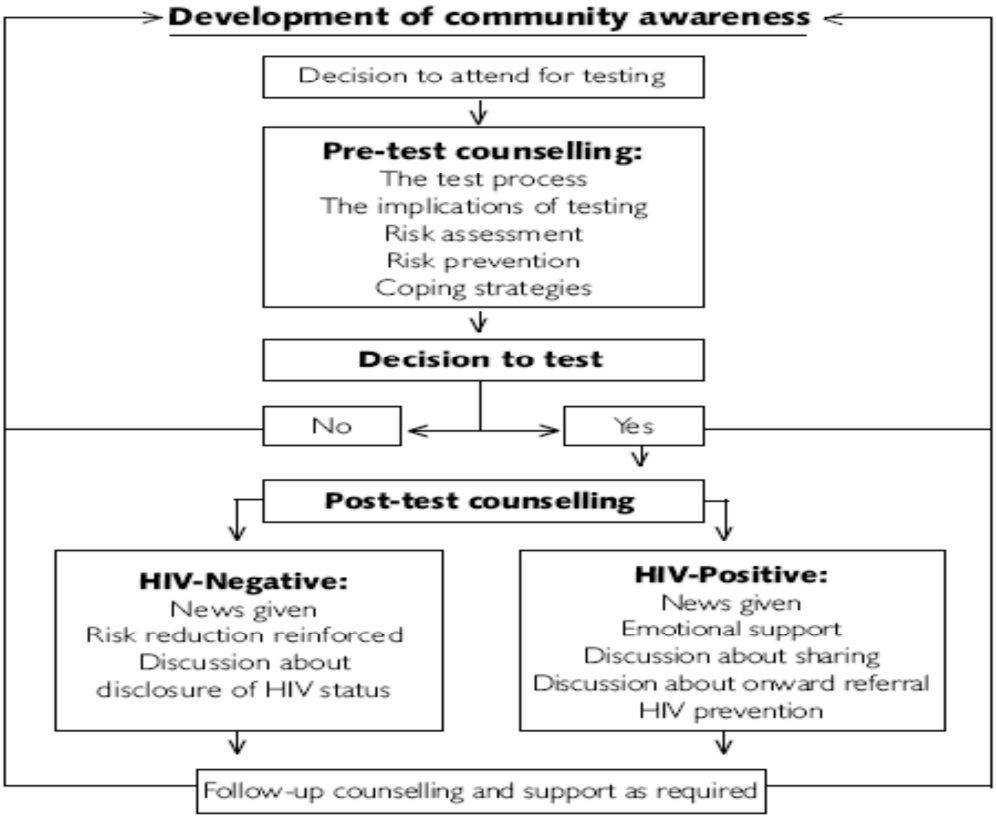


Figure 1. 1 Diagrammatic overview of the process of VCT Service

Source: (http://data.unaids.org/Publications/irc-pub01/jc379-vct_en.pdf)

As shown in the above diagram, the VCT process consists of pretest, post-test and follow-up counselling. HIV counselling can be adapted to the needs of the client/s and can be for individuals, couples, families and children and should be adapted to the needs and capacities of the settings in which it is to be delivered (UNAIDSHIV, 2000).

The content and method may vary considerably for men and women and with various groups, such as counselling for young people, men who have sex with men (MSM), injecting drug users (IDUs) or sex workers. The manner in which understanding of HIV sero status is very important in facilitating adjustment to impact of HIV infection.

Counselling as part of VCT ideally involves at least two major sessions (pretest counselling and post-test counselling). More time can be offered before or after the test, or during the period the client is coming up for test results.

Pre-test counselling: It should be offered before taking an HIV test. In an ideal world, the counsellor prepares the client for the test by explaining what an HIV test is, as well as by correcting myths and misinformation about HIV/AIDS. They may also discuss the client's personal risk profile, including discussions of sexuality, relationships, possible sex and/or drug-related behavior that increase risk of infection, and HIV prevention methods. The counsellor discusses the implications of knowing one's sero status, and ways to cope with that new information.

People who do not want pre-test counselling should not be prevented from taking a voluntary HIV test. However, informed consensus from the person being tested is usually a minimum ethical requirement before an HIV test.

Post-test counselling: The main goal of this counselling stage is very crucial to help clients understand their test results and initiate adaptation to their seropositive or negative status. It should always be offered.

When the test is seropositive, the counsellor tells the client the result clearly and sensitively, providing emotional support and discussing how he/she cope. During this session the counsellor must ensure that the person has immediate emotional support from a partner, relative or friend. When the client is ready, s/he may offer information on referral services that may help clients accept their HIV status and adopt a positive outlook.

In other way round, Counselling is also important when the test result is negative. While the client is likely to feel relief, the counsellor must emphasize several points. Here, discussion about how to changes in behavior that can help the client stay HIV-negative, such as safer sex practices including condom use and other methods of risk reduction. S/he also motivates the client to adopt and sustain new, safer practices and provide encouragement for these behavior changes.

During the “window period” (approximately 4-6 weeks immediately after a person is infected), antibodies to HIV are not always detectable. Thus, a negative result received during this time may not mean the client is definitely uninfected, and the client should consider taking the test again in 1-3 months.

2.4.2. Benefit of VCT Service

According to Mugula et al. (1995) VCT has a vital role to play within a comprehensive range of measures for HIV/AIDS prevention and support, and should be encouraged. It is an efficient and cost-effective strategy in expanding access to prevention, treatment and care services.

The potential benefits of voluntary counselling and testing service (FHI, 2004; Mugula et al., 1995) includes: facilitates behavior change, reduces the stigma attached to those who live with HIV/AIDS (Kathy and Stacey, 2009) and early VCT can lead to a delay in HIV deaths, better ability to cope with HIV related anxiety; awareness of safer options for reproduction and infant feeding; and motivation to initiate or maintain safer sexual and drug related behaviors. Other benefits include safer blood donation

World Health Organization (WHOS, 2005) summarized the benefits of knowing HIV status in three different levels such as at the individual, community and population.

- **At the individual**: VCT enhanced ability to reduce the risk of acquiring or transmitting HIV; access to HIV care, treatment and support; and protection of unborn infants.
- **At the community**: it creates a wider knowledge of HIV status and its links to interventions can lead to a reduction in denial, stigma and discrimination and to collective responsibility and action.
- **At the population level**: knowledge of HIV epidemiological trends can influence the policy environment, normalize HIV/AIDS and reduce stigma and discrimination.

2.5. The prevalence of HIV/AIDS in Ethiopian Higher education

Study show that the highest concentration of HIV positive people is between the ages of 15-40 (MOH, 2000). Besides, scholars in many part of the world point out that because of the high

concentration of people in this economically significant segment of the population of labor force age, productivity has been and continues to be negatively impacted by the factor such as HIV-related absenteeism, disability and the early death of experienced workers (Saint,2005).

Producing high level human resource to run an entire economy has relied on higher education. As a result society is undertaking and performing task effectively and efficiently based on the task of the organization missions.

Furthermore, HIV/AIDS holds the potential to undermine the country's substantial investments in education. It affects all people all over the world without discrepancy in race, religion and gender. Overall, it has major impact on distribution of education.

AIDS now exists within all regions of Ethiopia. The estimated national HIV infection rate is 2.1% (UNAIDS, 2007). This rate is roughly 1 in 50 Ethiopian adults aged range from 15 to 30 has the HIV virus. The highest at risk for HIV infection are female sex workers, children born from HIV-infected parents, and injection drug users.

Higher education communities are particularly vulnerable to HIV/AIDS due to their age groups. This vulnerability introduces a sizeable risk to the expected returns on investments made by families and government in the education of university or college students (Otaala, 2003). In spite of this risk, universities in Ethiopia have not established institutional policies or programs for the management and prevention of HIV/AIDS until 2005 (Saint, 2005). Institutional policies for the management of HIV/AIDS plays significant role for reduction of HIV infection.

This policy may include review of regulations on sick leave, confidentiality and the rights of persons living with AIDS, student counseling services to awareness programs, and curriculum content on testing facilities (Otaala, 2003).

According to Saint (2005) little has been known about the status of AIDS on Ethiopian University. As indicated by recent study in Jimma university which estimated that 12.2 % of the university communities were HIV positive (MOE, 2004). Since then Ministry of Education and university officials are discussed on issue to react against HIV infection in higher education.

As the result of this Addis Ababa university (AAU) became the pioneer that started the first HIV/AIDS policy in Ethiopia higher education at 2006 (AAU,2006). The policy document is established with a collaboration of African AIDS Initiatives International (AAII), but it started its VCT service in AAU in 2005.

According to the report of Addis Ababa highlight (AAH, 2006), the HIV/AIDS policy document served as a base and policy for other institutions in Ethiopia and out of Ethiopia, in the hope that other institutions of higher learning would follow in the footsteps of Addis Ababa University for the benefit of their young people.

As the oldest institution of higher learning in Ethiopia, Addis Ababa University played the leading role in promoting instruction, training and research in the fight against HIV epidemic.

The current student population of Addis Ababa University enrolled in the graduate, undergraduate, extension, and summer programs adds up to over 50,000. At present, AAU has nearly 2000 full time academic staff and over 5000 administrative personnel (AAUa, 2011).

The university is committed to playing active role in the fight against HIV/AIDS primarily by engaging in research in biomedical and psychosocial spheres. In doing this, it needs to establish collaborative relationships with scientific institutions both nationally and abroad. It also needs to work closely with national, regional and international partners in mitigating the impact of HIV/AIDS on its own staff and students as well as the wider Ethiopian society.

In this regards, the university encourage faculty, staff and students to submit to voluntary counseling and testing (VCT) and facilitate this by establishing the necessary facilities on campus. After establishment of a formal HIV policy document, various non-governmental organizations has involved in ant-AIDS movement. In this respect, an African AIDS initiatives International take the line share against mitigating HIV infection in higher education particularly in AAU.

As introduced in chapter one background of the study 1.1, AAII provides HIV prevention related service for the university community. According to Data Manager of AAII, the organization has two primary VCT centers in university i.e. 6 kilo and 4 kilos. Besides, it provides a VCT service outside the university community.

The organization collects and organizes client's record provided in each site on daily base. As the researcher observed the database of Epi Info software of the organization holds an estimated of 23,000 clients are visited the AAII VCT center both in campus and off campus site.

The AAII VCT center provides six days per week and 24 days per month services for university community. Its major activities include provision of quality VCT service; implementing quality information management system; referral network system establishment and monitoring feedback; strong VCT HIV sensitization and promotion; and replication of best practices to other centers.

Currently, the organization extends its VCT service from AAU to other Ethiopian higher education including Dbere Marko's University and Mezan Tepi University.

Africa AIDS Initiative International Inc.

TOTAL NUMBER OF VCT CLIENTS COUNSELED AND TESTED

FROM MARCH 2005 - OCTOBER 31/2011

	Number of Clients Counseled			Number of Clients Counseled & Tested			Test Result						Post test Counseled		
							Sero-Positive			Sero-Negative					
Category	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T
Student	9652	4433	14085	9554	4402	13956	43	41	84	9511	4361	13872	9554	4402	13956
Staff	1078	816	1894	1073	809	1882	26	29	55	1047	780	1827	1073	809	1882
Community	4436	2771	7207	4404	2759	7163	199	266	465	4205	2493	6698	4404	2759	7163
Total	15166	8020	23186	15031	7970	23001	268	336	604	14763	7634	22397	15031	7970	23001

Table 2. 1 Annual report of AAII VCT Client from March 2005-October2011

The above table indicates that the number of student who has HIV positive test result is greater than from that of staff but less than community. Besides, the total number of client who checks their status in this range is better than other user of the center. According to this report the percentage of females (56%) are more exposed for HIV than male (44%).

The data in all centers are collected using the same standard paper format that are used by CDC (Center of Disease control and Prevention) and managed by VCT EPi Info software system.

As shown in table 2.2 the initial data are collected from AAII VCT centers for seven years. The total clients visited the centers up to October, 2011 are 23,000.

The data collected over time from client can tell so much if appropriately analyzed as to what factor predominantly cause HIV infection. However, when data is getting larger and larger, the data carry a lot of implicit knowledge that cannot be accessed with simple statistical analysis methods. Hence application of other tool is inevitable. So the most appropriate technique to discover and manage knowledge implicitly hidden inside the collected data is data mining (Abdule-kereem et al.,2000).

We discussed the importance of VCT service in section 2.4.2, the data that we obtain from this center is very crucial and play significant role in decision making related to HIV/AIDS control and prevention.

CHAPTER THREE

LITERATURE REVIEW

In this chapter, an attempt has been made to review the literature on the concepts and techniques of data mining in general and its application in the health care sector in particular which is aimed to provide background about the models to be built.

3.1. Overview of Data Mining

Across a wide variety of fields, data were being collected and accumulated at a dramatic pace. The rapid growth and integration of databases provides scientists, engineers , and business people with a vast new resources that can be analyzed to make scientific discoveries, optimize industrial systems and uncover valuable patterns **(**Hand et al ., 2001).

The developments in digital data acquisition and storage technology have resulted in the growth of huge databases. This has occurred in all area of human attempt , from the ordinary (such as supermarket transaction data, credit card usage records, telephone call detail, and government statistics) to the more exotic (such as images of astronomic bodies , molecular databases, and medical records)(Thearling et al. ,nd).The large size and complexity of such data in many scientific domains becomes impractical to manually analyze, explore, and understand data.

To undertake large data analysis projects, researchers and practitioners have adopted established algorithms from statistics , machine learning, neural networks ,and databases and have also developed new methods targeted at large data mining problems **(**Hand et al., 2001). Witten and Frank (2005)states that lack of data is no longer a problem at a current stage. However, the inability to generate useful information from data is the problem. On other hand, it was recognized that information is at the heart of business operation and that decision make use of the data stored to gain valuable insight in to the business (Rea, 1995). Thus there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data (Fayyad et al., 1996). These theories and tools are the subject of the emerging fields of knowledge discovery in database and data mining.

3.2. Knowledge Discovery in Database (KDD) and Data mining

Conventionally the idea of searching pertinent patterns in data has been referred using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely (Fayyad et al, 1996).

The term knowledge discovery in databases (KDD) is synonymously used with data mining. The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro, 2000). The purpose was to emphasize that knowledge is the end product of a data-driven discovery and it has been popularized in the 'artificial intelligence' and 'machine-learning' fields.

In Fayyad's et.al. (1996) view, KDD refers to the overall process of discovering useful knowledge from data and that data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

Trybula (1997) also states that knowledge discovery (KD) is the process of transforming data into previously unknown or unsuspected relationships that can be employed as predictors of future action. Furthermore, he asserts that KDD is a term that has been employed to encompass both data mining and KD. Essentially, the basic tasks of data mining and KD are to extract particular information from existing databases and convert it into understandable or sensible conclusions (i.e., knowledge).

3.3. What is Data mining?

In the Information Technology era information plays vital role in every sphere of the human aspects. Thus, to efficiently inspire information, it is very important to generate information from massive collection of data. The data can range from simple numerical figures and text documents to more complex information such as spatial data, multimedia data, and hypertext documents (Deshpande and Thakare, 2010). However, the huge size of these data sources makes it impossible for a human experts to come up with interesting information or patterns that help in the proactive decision making process. Therefore, to take complete advantage of data; the data retrieval is not enough. It requires a tool for automatic summarization of data, extraction of the essence of

information stored, and the discovery of patterns in raw data. This tool is called data mining (Deshpande and Thakare, 2010).

Data mining is one step in the KDD process (Fayyad et al., 1996). It is the most researched part of the process. Data mining can be used to verify a hypothesis or use a discovery method to find new patterns in the data that can do predictions with new, unseen data.

Many definitions of data mining could be found in the literature, Berry et.al. (1997) define it as the exploration and analysis of large quantities of data by automatic or semiautomatic means in order to discover meaningful patterns and rules. According to Bigus (1996), data mining is the efficient discovery of valuable, non-obvious information from a large collection of data.

Witten (2000) also defines as data mining is the analysis of large observational data sets with the goal of finding unsuspected relationships. A data set can be "large" either in the sense that it contains a large number of records or that a large number of variables is measured on each record.

Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data; The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases (Two Crows Corp ,1999); It is the process of finding mathematical patterns from (usually) large sets of data; these can be rules, affinities, correlations, trends, or prediction models (Berry and Linoff,2000).

Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (WKDD ,2008).

3. 4. Knowledge Discovery Process Models (KDPM)

The knowledge discovery process can be formalized into a common framework using process models. The main reason for introducing process models is to formalize knowledge discovery projects within a common framework, a goal that result in cost and time savings, and improve understanding, success rates, and acceptance of such projects. The models emphasize independence from specific applications, tools, and vendors (Cios et al., 2007).

A KD process model provides an overview of the life cycle of a data mining project, containing the corresponding phases of a project, their respective tasks, and relationships between these tasks. There are different process models originating either from academic or industrial environments. All the processes have in common the fact that they follow a sequence of steps which more or less resemble between models

Although the steps are executed in a predefined order and depend on one another (in the sense that the outcome of a step can be the input of the next step), it is possible that, after execution of a step, some previous steps are re-executed taking into consideration the knowledge gained in the current step (feedback loops).

An interesting issue that arises in any KDD process model is choosing which algorithm or method should be used in each phase of the process to get the best results from a dataset. Since taking the best decision depends on the properties of the dataset, Popa et al. (2007) proposes a multi-agent architecture for recommendation and execution of knowledge discovery processes.

Knowledge discovery in database (KDD) and data mining is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the internet and the widespread use of databases have created an immense need for KDD methodologies. In this paper, we provide an overview of common knowledge discovery process model and approaches to solve tasks. The challenge of extracting knowledge from data draws upon research in statistics and database (Mamcenko and Beleviciute, 2007).

A as result, data mining research require stable and well defined foundation which are well understood and popularize thought the community (Kurgan &Musilek,2006). The primary objective of data mining process or methodologies is building stable models following some logical steps (Berry &linoff,2004) .

According to Santos & Azevedo (2008) the followings are commonly used Data mining process model / methodology /KDPM that data mining appliedin research; Knowledge Discovery in Database (KDD), CRoss-Industry Standard Process for Data Mining(CRISP-DM), Sample, Explore, Modify, Model, Assess(SEMMA) and Hybrid model.

3.4.1. KDD Process Model

The essence of Academic research model was started in the mid-1990s, when the DM field was being shaped; researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain (Cios et al., 2007). As a result different Academic research models are developed and followed by industry. According to Fayyad et al. (1996). KDD process model was the leading Academic research model applied in different data mining project.

KDD process, as presented in (Fayyad et al, 1996), is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are five stages as presented in figure 3.1:

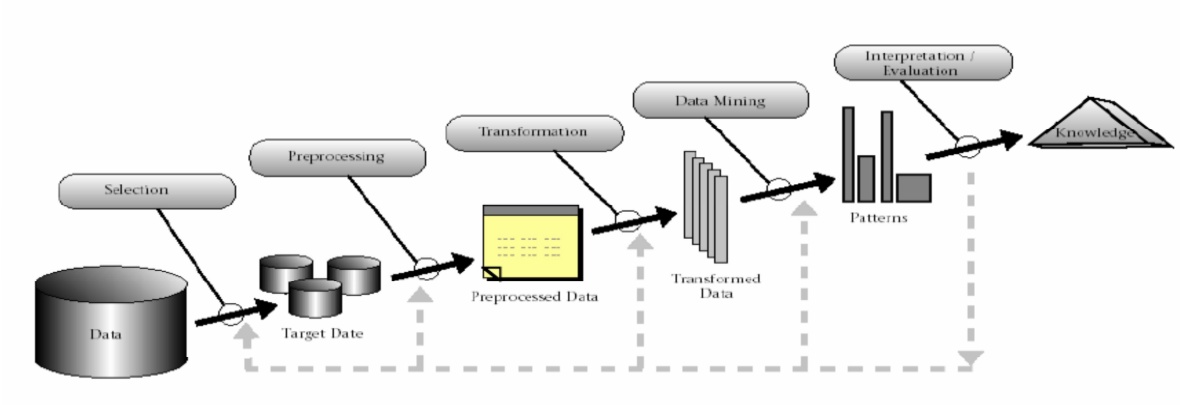


Figure 3. 1 Overview of KDD

An outline of the steps in Figure 3.1 has been adequate for understanding the concepts required for the KDD process. The following are the steps involved:

Selection - this stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed;

Pre-processing - this stage consists on the target data cleaning and preprocessing in order to obtain consistent data. Here we also try to eliminate noise that is present in the data. Noise can be defined as some form of error within the data. Some of the tools used here can be used for filling missing values and elimination of duplicates in the database.

Transformation - this stage consists on the transformation of the data using dimensionality reduction or transformation methods. Usually there are cases where there are a high number of attributes in the database for a particular case. With the reduction of dimensionality we increase the efficiency of the data-mining step with respect to the accuracy and time utilization.

Data Mining - The data mining step is the major stage in data KDD. This is when the cleaned and preprocessed data is sent into the intelligent algorithms for classification, clustering, similarity search within the data, and so on. Here we chose the algorithms that are suitable for discovering patterns in the data. Some of the algorithms provide better accuracy in terms of knowledge discovery than others. Thus selecting the right algorithms can be crucial at this point.

Interpretation/Evaluation –in this stage the mined data is presented to the end user in a Human-viewable format. This involves data visualization, which the user interprets and understands the discovered knowledge obtained by the algorithms.

3.4.2. CRISP-DM

A second most popular process model that has proven application is Cross Industry Standard Process Data Mining (CRISP-DM). CRISP-DM is a product neutral data mining model developed by consortium of several companies (Roiger and Geatz, 2003). It is a standard process model in industries which consisting of a sequence of steps that are usually involved in a data mining study .CRISP-DM was developed by the means of the efforts of an association initially composed with Daimler Chrysler, SPSS and NCR (Chapman et al.,2000). The life cycle of a data mining project consists of six phases. These are business understanding, data understanding, data preparation, modelling, evaluation and deployment .CRISP-DM process consists of six phases (Roiger and Geatz, 2003).

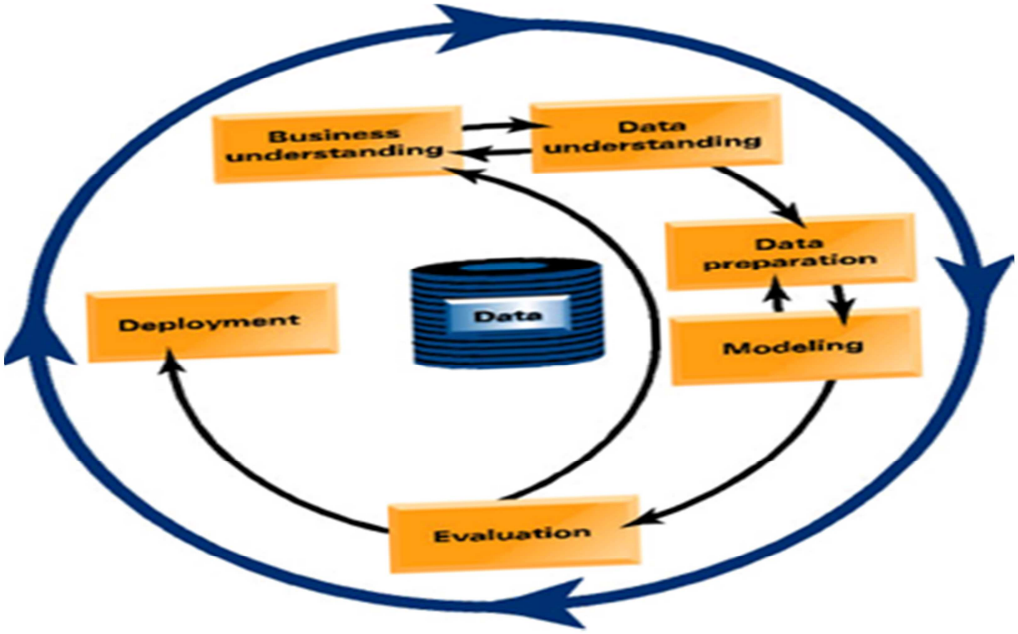


Figure 3. 2 CRISP-DM process model adapted from (Daniele and Satheesh ,n.d.)

CRISP-DM is complete and well documented. All this stages are duly organized, structured and defined, allowing that a project could be easily understood or revised (Santos & Azevedo, 2005).

As seen from the figure 3.2, CRISP-DM model is divided into six phases and related tasks as described in the [Daniele and Satheesh ,(n.d.); Chapman et al.,2000] as follows:

1. **Business understanding**: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives [Daniele and Satheesh ,(n.d.)]. Some of the objectives of this phase are determine business objectives and business success criteria assess situation such as inventory resource requirements, assumptions constraints risk contingencies cost and benefits determine data mining goals produce project plan initial assessment of tools and techniques. Business understanding is critical because it identifies the business objectives and, thus, the success criteria of data mining projects (Koh, and Tan, 2005).

2. Data understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. The major activities are collect initial data (initial data collection report), describe data (data description report), explore data (data exploration report) and verify data quality (data quality report).

3. Data preparation: The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order. Data preparation and cleaning is an often neglected but extremely important step in the data mining process models. Often, the method by which the data where gathered was not tightly controlled, and so the data may contain outlier values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes). Analyzing data that has not been carefully screened for such problems can produce highly misleading results, in particular in predictive data mining [Daniele M. and Satheesh R,.nd)].

The main tasks in data preparation are; select data which is rationale for industry exclusion or inclusion clean the data and document it, construct data derived attributes and generated reports integrate and merge data, format data reformat data and have dataset description. Data preparation step is by far the most time-consuming part of the knowledge discovery process (Cios et al., 2007).

4. Building Model: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phases such as select modeling techniques and modeling assumptions, generate test design, build model, parameter settings and model description and assess model, model assessment, revise parameter settings.

5. Evaluation: Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives. At the end of this phase, a decision should be reached on how to use the DM results through assessment of data mining results with respect to business success criteria to

approve models, review of process, and determine next steps such as possible actions and decisions making based on patterns and relationships.

6. Deployment: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained need to be organized and presented in a way that the end user can use and understand it. The main activities in this phase are plan deployment, plan monitoring and maintenance (monitoring and maintenance plan), produce final plan and present, review project and produce documentation. According to Chapman et al. (2000), depending on the requirement, these phase can be as simple as generating a report as complex as implementing a repeatable data mining process across the enterprise.

According to kurgan and Musilek (2006) the CRISP-DM methodology is the most suitable process model for novice data miners due to its easy, complete documentation , and intuitive , industry-applications-focused description.

The CRISP-DM model has been used in domains such as medicine, engineering, marketing, and sales. It has also been incorporated into a commercial knowledge discovery system called Clementine (see SPSS Inc. at http://www.spss.com/clement and Cios et al., 2007).

3.4.3. The SEMMA Process

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA stands for(**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess) refers to a methodology that clarifies this process (Obenshain, 2004). The SAS Institute considers a cycle with five stages for the process:

Sample – this stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.

Explore - this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify - this stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model - this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals (Santos & Azevedo, 2005).

3.4.4 Hybrid model

The development of both academic particularly the KDD and industrial oriented (CRISP-DM and other) data mining models has led to the growth of hybrid models, i.e., models that combine the features and job of both. Hybrid model is a six-step KDP model (see Figure 3.3) developed by Cios et al.(2007). It was developed mainly based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include introducing several new explicit feedback mechanisms and in last steps the knowledge discovered for a particular domain may be applied in other domains.

According to Cios et al.(2007) the description of the six steps are explained below.

Step 1: understanding of the problem domain. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. The major tasks undertaken in this phase are: Understand and learn domain-specific terminology.

- explain the problem area in detail and make a boundary for it
- Lastly, convert the project goals into DM goals, and select appropriate data mining tools to be used in performing the whole research.

Step 2: understanding of the data. In this phase collecting sample data and deciding the types, format and size of data was a primary task. Background knowledge can be used to guide these efforts. Besides, data should be checked for completeness, redundancy, missing values,

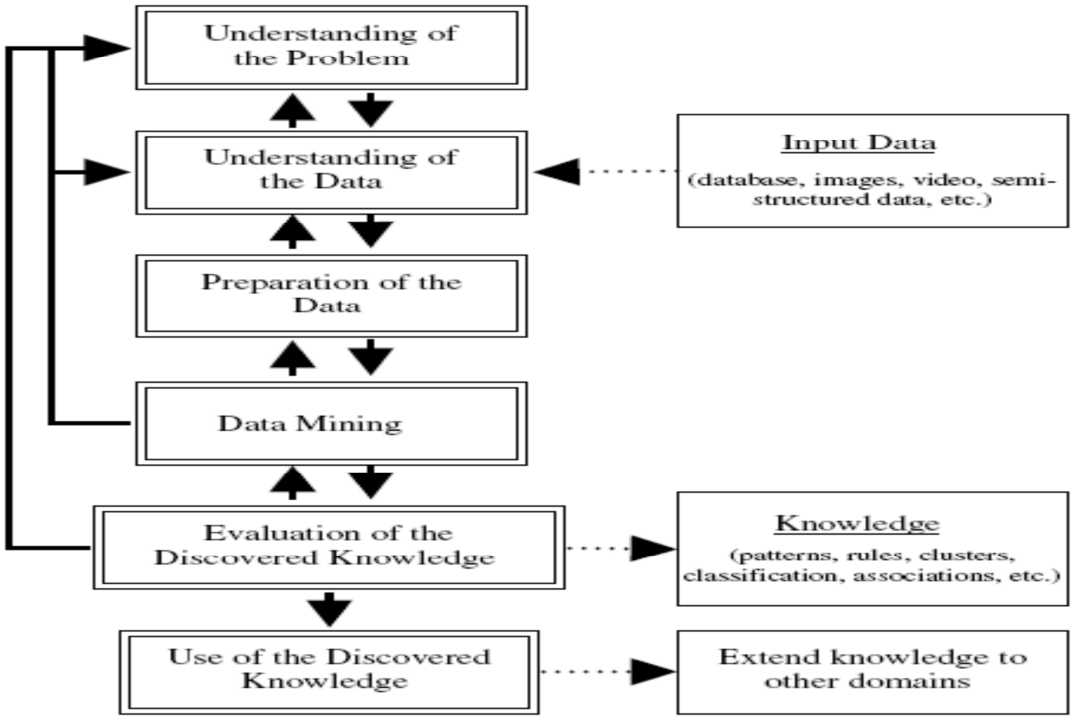


Figure 3. 3 The six phases of Hybrid data mining model(Pal, et al.,2005)

plausibility of attribute values, etc. verification of the usefulness of the data with respect to the DM goals was also a major activity undertaken in this step.

Step3: Preparation of the data. Once the problem is understand and selected complete and appropriate data to the problem under investigation the next step be selecting representative data. As a result, the researcher be used only the selected data as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization).

The end results are data that meet the specific input requirements for the DM tools selected in Step 1.This steps is the most time consuming part of all data mining model (Cios et al., 2007).

Step4 Data mining: Another key step in the knowledge discovery process is data mining. It involves the use of several DM tools on data prepared in step 3.The application of this tools discover new knowledge. The process of discovering new information includes: the data model was constructed using one of the chosen DM tools and training and testing procedures are designed. Then generated data model was verified by using testing procedures. It takes less time than data preprocessing steps. Some of the common data mining technique implemented using data mining tools are decision tree, neural networks, clustering, support vector machine, rule induction, association rule.

Step 5 Evaluation of the discovered knowledge. Evaluation includes understanding the results, checking whether the discovered knowledge was novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process was revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is also prepared.

Step 6 Use of the discovered knowledge. The final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge was created and the entire project will be documented and reported.

3.4.5. Comparison of SEMMA, KDD and CRISP-DM

Today, research efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made. The subsequent models provide more generic and appropriate descriptions. Most of them are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application (kurgan and Musilek ,2006). Santos & Azevedo (2008) have summarized the association of the three most popular process model steps as shown in Table 2.1

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

Table 3. 1 Comparison of Data mining process model

As shown in table 3.1 most of the steps to be followed in all methods look like similar. Nevertheless, KDD and SEMMA do not have a step "understanding the problem" before selection and sample steps respectively. Undoubtedly, understanding the domain problem is often a prerequisite steps but missed by the three methodology. Considering the presented analysis we conclude that SEMMA and CRISP-DM can be viewed as an implementation of the KDD process described by (Fayyad et al, 1996).

In general, According to Refaat (2007) all methodologies contain and followed same data mining tasks. Hence this research be conducted based on understanding of the domain problem and business context, consequently, the best fit data mining model to conduct the research is inevitably hybrid model. This model completes the above missed limitation in methodology.

3.5. Data Mining Task

The goal of data mining is to **learn** from data, and there are two broad categories of data mining strategies: supervised and unsupervised learning. Supervised learning methods are deployed (or strategically put into service in an information technology context) when values of variables (inputs) are used to make predictions about another variable (target) with known values. Unsupervised learning methods can be used in similar situations, but are more frequently deployed on data for which a target with known values does not exist (Obenshain, 2004).

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and

predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions (Han and Kamber, 2006; Kantardzic, 2003). According to Fayyad et.al. (1996), the data mining stage of KDD process consists of two main types of data mining methods: verification oriented (the system verifies user's hypothesis) and discovery oriented (the system finds new rules and patterns autonomously). The discovery-oriented methods can be further partitioned into descriptive (e.g. Visualization) and predictive (like regression and classification).

In predictive modeling tasks, one identifies patterns found in the data to predict future values (Levin and Zahavi, 1999). Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. Whereas descriptive techniques are sometimes referred to as unsupervised learning since there is no already-known result to guide the algorithms (Two crows corporation, 1999).

On the other hand, descriptive models belong to the realm of unsupervised learning. Such models interrogate the database to identify patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, among others, belong to this family of descriptive models (Levin and Zahavi, 1999; Han and Kamber, 2001).

The Predictive modeling is based on the use of the historical data. Predictive model data mining tasks include classification, regression, time series, and prediction. Unlike the predictive modeling approach, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties (Dunham, 2003). Clustering, summarization, association rules, and sequence discovery are viewed as descriptive in nature (shown in Figure 3.3: Dunham, 2003).

3.5.1. Classification and Prediction

Classification is one of the most common data mining tasks, which is also pervasive in human life. Classification finds common properties among different entities in order to organize them into predefined classes (Han et al. ,2001). Human beings usually classify or categorize in order to understand and communicate about the world. For any object or instance, classes are predefined according to the value of a specific field (Berry and Linoff, 1997).

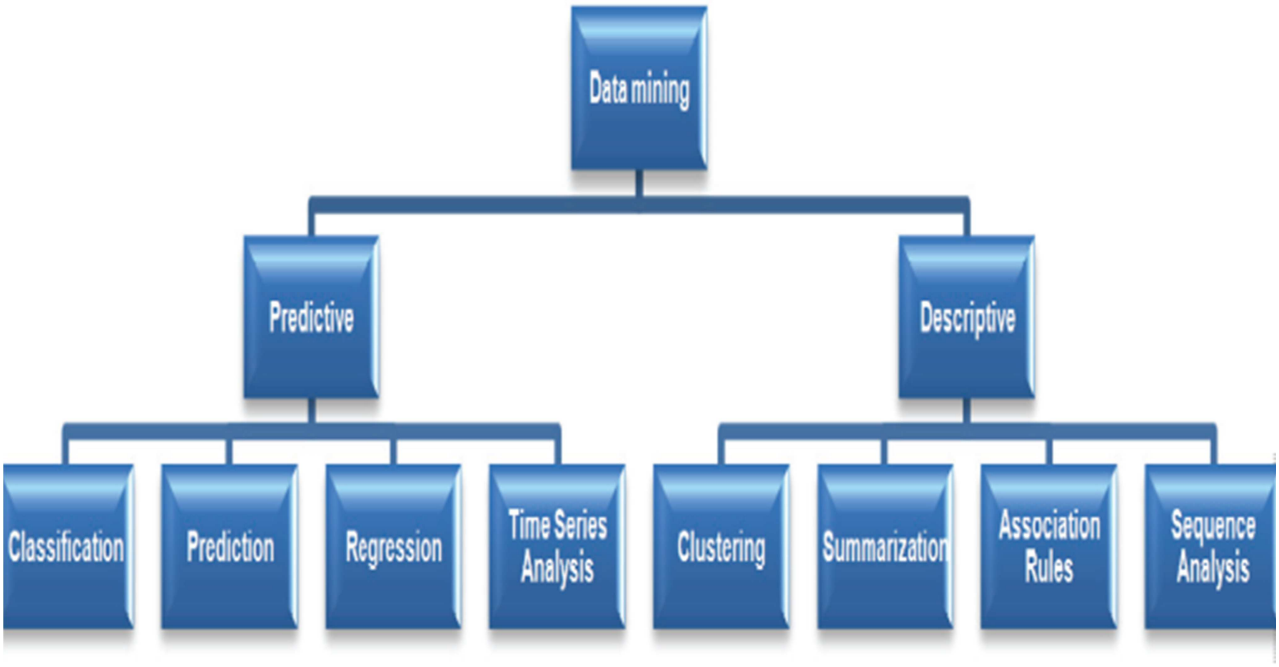


Figure 3. 4 Classification of Data mining models and tasks (Dunham, 2003)

There are many different methods, which may be used to predict the appropriate class for the objects or stations. Among the most popular ones are: logistic regression, decision tree, rule induction, case-based reasoning, and neural network (Moshkovich et al., 2002).

Even though , there are many data mining tasks available and commonly used in various application, the researcher used decision tree and rule induction techniques in this research because their result is simple to explain for end user and these technique support the selected data mining tasks for this research. The most widely used techniques for classification are decision trees, neural networks and memory-based reasoning (Berry et al., 1997).

3.5.1.1. Decision Trees

One of the most commonly used data mining techniques for classification tasks are decision trees. It is a simple knowledge representation technique and classifies instances in to a finite number of classes (Larvac, 1998). In decision tree induction, the nodes of the tree are labeled with attribute names, the edges of the tree are labeled with possible values for the attributes and the leaves of the tree generate decision tree from a given set of attribute-value tuples. A decision tree is constructed

by repeatedly causing a tree construction algorithm in each generated node of the tree (Larvac, 1998 ; Two crows corporation,1999).

The classification is performed separately for each leaf, which represents a conjunction of attribute valued in a rule (Last and Kandel, 2002). Structurally, decision trees consist of two types of nodes; non-terminal (intermediate) and terminal (leaf). The former correspond to questions asked about the characteristic features of the diagnosed case. Terminal nodes, on the other hand generate a decision (Rudolfer et al., 2002). Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that be used to make predictions.

An important quality of decision trees is that they handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets (Two crows corporation, 1999). Rogers and Joyner (2002) also stated "tree based models are good at selecting important variables, and therefore work well when many of the predictors are irrelevant."

A common criticism of decision trees is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In addition, all splits are made sequentially, so each split is dependent on its predecessor (Two crows corporation, 1999).

3.5.1.1.1. Decision tree Construction Methods

All decision tree construction methods are based on the principles of recursively partitioning the dataset until homogeneity is achieved. The construction of decision tree involves the following three phases (Sheikh, 2008):

1) Construction phases: the initial decision tree is constructed in these phases, based on the entire training dataset. It require recursively partitioning the training set in to two or more sub- partitions using a splitting criterion, until a stopping criteria is met. The basic strategy of construction phase is described as follows. The tree starts as a single node representing the training sample. If the samples are of the same class, then the node become a leaf and is labeled with that class. Otherwise, the algorithms uses entropy-based measure known as information

gains as a heuristic for selecting the attribute that best separate the sample s into individuals classes. The information gain measure is used to select the test attribute at each node in the tree. The attribute with the highest information gain is chosen as the test attribute for the current node.

2) Pruning phase: The pruning phase involves removing some of the lower branches and nodes to improve performance.

3) Processing the pruning tree: In this stage decision tree is proposed to improve understandability. Various listed decision tree algorithms such as CHAID, C4.5/C5.0, CART, ID3, J48 and many other with less familiar algorithms produce tree that differ from one another in the number of splits allowed at each level of tree , how those splits are chosen when the tree is built(Thearling et al. , n.d).

3.5.1.1.2. Decision tree Algorithms

The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items (Rokach and Maimon, 2005). A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), Quest, and C5.0,J48 (Two crows corporation, 1999).

J48 decision tree algorithms adopt an approach in which decision tree models are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built (Han & Kamber, 2006). According to Bharti et al. (2010), J48 decision tree algorithm is a predictive machine learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. It can be applied on discrete data, continuous or categorical data. This algorithms was used in this study to classify the VCT clients as HIV infected or non-HIV infected. The J48 decision tree can serve as a model for classification as it generates simpler rules and remove irrelevant attributes at a stage prior to tree induction. In

several cases, it was seen that j48 decision trees had a higher accuracy than other algorithms (Witten and Frank 2000). It offers also a fast and powerful way to express structures in data

To view the internal operation of the J48 algorithms Han and kamber (2006) explained how the algorithms works in detail step by step as follows.

Figure 3. 5 Basic pseudo code for J48 decision tree algorithms

Algorithm: Generate decision tree.

Input: Sets of training dataset (D), Attribute list ,Attribute selection method;

Output: A decision tree.

Procedures:(1) Create a node N; (2) If tuples in D are all of the same class, C then

(3) Return N as a leaf node labeled with the class C;

(4) If attribute list is empty then

(5) Return N as a leaf node labeled with the majority class in D; // majority voting

(6) Apply Attribute selection method (D, attribute list) to find the "best" splitting criterion;

(7) Label node N with splitting criterion;

(8) If splitting attribute is discrete-valued and multi way splits allowed then // not restricted to binary trees

(9) Attribute list ← attribute list- splitting_ attribute; // remove splitting attribute

(10) For each outcome j of splitting criterion // partition the tuples and grow subtrees for each partition

(11) Let Dj be the set of data tuples in D satisfying outcome j; // a partition

(12) If Dj is empty then (13) attach a leaf labeled with the majority class in D to node N;

(14) Else attach the node returned by Generate decision tree (Dj, attribute list) to node N; Endfor
(15) Return N;

In general, according to Han and Kamber (2006) the J48 decision tree algorithms follows five main procedures including:

1. Choose an attribute that best differentiates the output attribute values.

2. Create a separate tree branch for each value of the chosen attribute.

3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.

4. For each subgroup, terminate the attribute selection process if:

a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.

b. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.

5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

The algorithm is applied to the training data. The created decision tree is tested on a test dataset, if available. If test data is not available, J48 performs a cross-validation using the training data itself.

3.5.1.2. Neural Networks

A Neural Network is an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning (Rea, 2001). Neural networks are developed due to difficulty in creating the basic intelligence in computers using the conventional algorithmic approach (Pudi,2001). According to Berry and Linoff (1997), Bigus (1996) and Atiezenza et. al. (2002) Neural Networks are also probably the most common data mining classification.

Neural networks (NN) are those systems modeled based on the working of the human brain. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it.

The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input (Gupta et al., 2011).

These networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

3.5.1.3. Rule Induction

Rule induction is one of the techniques most used to extract knowledge from data, since the representation of knowledge as if/then rules is very intuitive and easily understandable by problem-domain experts (Wong and Leung, 2000). It is an area of machine learning in which formal rules are extracted from a set of observations. The extracted rules may represent a full model of the data (in the form of a rule set) like in Clark and Niblett (1989), or represent local patterns in the data (in the form of individual rules) like in Agrawal et al. (1996). Hence, regularities hidden in data are frequently expressed in terms of rules , rule induction is one of the fundamental tools of data mining at the same time (Grzymala-busse, n.d.). Usually rules are expressions of the form

if (attribute − 1; value − 1) and (attribute − 2; value − 2) and ……and (attribute − n; value − n) then (decision; value):

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain(Grzymala-busse, nd.). Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labeled as attributes and a decision. We restrict our attention to rule induction which belongs to supervised learning: all cases are pre classified by an expert (Grzymala-busse, n.d.). In other words, the decision value is assigned by an expert to each case. Attributes are independent variables and the decision is a dependent variable. Rule induction sometimes called rule learner is also perhaps the form of data mining that most closely resembles the process that most people think about when they think about data mining, namely “mining” for gold through a vast database. The gold in this case refers to the

rule that is interesting- that tell you something about your database that you did not already know and probably weren't able to explicitly articulate(Thearling et al. ,n.d).

Rule induction on a database is a process undertaking using intelligent software where all possible patterns are systematically pulled out of the data . In this process the accuracy is added to them that tell the user how strong the pattern is and how likely it is to occur again (Thearling et al. ,n.d). It has been widely used to represent knowledge in expert system and they have the advantage of being easily interpreted by human experts because of their modularity(Cabena etal.,1998; Moshkovich et al, 2002). Rule induction systems are highly automated and are probably the best of data mining techniques for exposing all possible predictive patterns in a database. They can be used in prediction problem but algorithm for combining evidence from a variety of rules come from practical experience (Thearling et al. ,n.d)

In rule induction, rules are expressed in the form of "if this, this and this then this". To make the rules useful, we require two important information including **accuracy** and **coverage.** This because the pattern in the database is expressed as rule does not mean that it is true all the time. Similar to other data mining algorithms it is important to recognize and make explicit the uncertainty in the rule, i.e. how often is the rule correct. This is what the "accuracy" of the rule means. The coverage of the rule has to do with how much of the database the rule "cover" or how often the rule applies (Thearling et al. ,n.d). Summary of coverage and accuracy be showed below.

	Accuracy Low	Accuracy High
Coverage High	Rule is rarely correct but can be used often.	Rule is often correct and can be used often.
Coverage Low	Rule is rarely correct and can only be used rarely.	Rule is often correct but can only be used rarely.

Table 3. 2 Rule coverage versus accuracy adapted from (Thearling et al. ,nd).

3.5.1.3.1. Rule induction for Prediction

The objective of data mining is to extract valuable information from one's data, to discover the 'hidden gold'. In Decision Support Management terminology, data mining can be defined as 'a decision support process in which one search for patterns of information in data' (Parsaye, 1997).

Data mining techniques are based on data retention and data distillation. Rule induction models belong to the logical, pattern distillation based approaches of data mining. These technologies extract patterns from data set and use them for various purposes, such as prediction of the value of a dependent field (Matsatsinis et al., n.d.).

After the rules are created and their interestingness is measured there is also a call for performing prediction with the rules. Each rule by itself can perform prediction; the consequent is the target and the accuracy of the rule is the accuracy of the prediction. But because rule induction systems produce many rules for a given antecedent or consequent there can be conflicting predictions with different accuracies. This is an opportunity for improving the overall performance of the systems by combining the rules. This can be done in a variety of ways by summing the accuracies as if they were weights or just by taking the prediction of the rule with the maximum accuracy (Thearling et al. ,n.d).

3.5.1.3.2. Rule Induction Algorithms

An induction algorithm takes as input specific instances and produces a model that generalizes beyond these instances. There are two major approaches by induction algorithm: supervised and unsupervised. In the supervised approach, specific examples of a target concept are given, and the goal is to learn how to recognize members of the class using the description attributes. In the contrary, unsupervised approach is provided with a set of examples without any prior classification, and the goal is to discover underlying regularities or patterns by identifying clusters or subsets of similar examples (My Data Mining Weblog (MDMW), 2008).

This research is deal with supervised induction algorithm methods because of simplicity and the goal of the research is prediction about the labeled or known class.

Rule induction seeks to go from the bottom up and collect all possible patterns that are interesting and then later use those patterns for some prediction target. It also retains all possible patterns even if they are redundant or do not aid in predictive accuracy. Hence, in a rule induction system if there were two columns of data that were highly correlated (or in fact just simple transformations of each other) they would result in two rules (Thearling et al., n.d).

Rule induction is also known as **Separate-And-Conquer method**. This method apply an iterative process consisting of first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process (MDMW, 2008). Some examples of these kinds of systems which are supported by WEKA software are discussed below:

1) **OneR:** OneR or "One Rule" is a simple algorithm proposed by Holt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its 'one rule'. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random (Holte,1993).

2) **PART:** PART is a separate-and-conquer rule learner proposed by Eibe and Witten (2005). The algorithm producing sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning (Frank and Witten, 1998)

3) **Decision Table:** It is the method used to build a complete set of test cases without using the internal structure of the program in question. In order to create test cases we use a table to contain the input and output values of a program. It summarizes the dataset with a 'decision table' which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table (Kohavi,1995).

3.5.1.4. Support Vector Machine

A support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis (Vapnik ,1995.) SVM was first proposed and developed by Vapnik (1995).It was gained popularity Piroozniam and Deng (2006) due to many promising features such as:

- Better empirical performance than other classification technique
- The algorithm has a strong theoretical foundation, based on the ideas of VC (Vapnik Chervonenkis) dimension and structural risk minimization
- **Scale up:** the SVM algorithm scales well to relatively large datasets
- **Robustness:** the SVM algorithm is flexible due in part to the robustness of the algorithm itself, and in part to the parameterization of the SVM via a broad class of functions, called kernel functions. The behavior of the SVM can be modified to incorporate prior knowledge of a classification task simply by modifying the underlying kernel function and
- **Accuracy** the most important explanation for the popularity of the SVM algorithm is its accuracy. The underlying theory suggests an explanation for the SVMs excellent learning performance, its widespread application is due in large part to the empirical success the algorithm has achieved.

Support Vector Machine is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. The algorithm performs discriminative classification, learning by example to predict the classifications of previously unclassified data (Piroozniam and Deng, 2006).

The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

3.5.1.4. 1. Method of SVM Classification

The original SVM classification attempts to find the best separating hyper plane with the largest margin width and the smallest number of training errors, as depicted in Figure 3.6

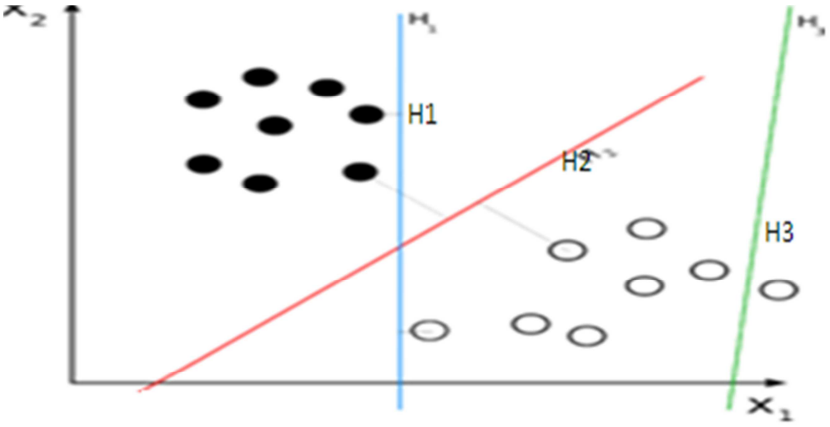


Figure 3. 6 Visualization of linearly separable problem with marginal line

As we can observe in the above figure H3 doesn't separate the two classes. H1 does, with a small margin and H2 with the maximum margin. Besides, it is easy to imagine that the hyper planes that can separate the training data perfectly, the one that has the maximum margin width has a better chance to give the best generalization performance on future unseen data. For this reason, SVM carry out the model of recognition for two-class problems by determining the hyper plane of separation with the maximum distance to the narrowest points of the positioning of formation. These points are called the vectors of support (Chang, et al., 2010).

The learning algorithm of SVM is figure out the boundary that separates the two different classes of training data, subject to certain optimization criteria. Therefore, the exact profile of the boundary is determined by the training instances located in the proximity of the boundary. Consequently, those training instances that are far away from the boundary essentially play no role in determining the boundary. In SVM such a hyper plane with maximum margin is determined by solving a convex quadratic programming (QP) problem. In this section, we present from a mathematical point of view, only the model applied in this study, as the case where the data are linearly separable. According to Han and kamber (2006) discussion the algorithms of Linear SVM work for linearly separable data sets.

Let the data set D be given as (X1, y1), (X2, y2)…, (X |D|, y|D|), where Xi is the set of training tuples with associated class labels , yi.

Each yi can take one of two values, either +1 or-1 (i.e., yi ϵ (-1, +1), for this research case corresponding to the classes HIV = -Ve and HIV =+VE respectively.

There are an infinite number of separating lines that could be drawn to separate all of the tuples of class +1 from all of the tuples of class-1.However, finding the "best" one, that is, one that (we hope) have the minimum classification error on previously unseen tuples is under research.

To solve the stated problem and to illustrate the separating line we use the term "hyper plane" to refer to the decision boundary that we are seeking, regardless of the number of input attributes. The application of SVM approaches this problem by searching for and suggesting for the maximum marginal hyper plane. Consider Figure 3.7 which shows two possible separating hyper planes and their associated margins.

There are an infinite number of (possible) separating hyper planes or "decision boundaries". According to the above figure 3.7 both hyper planes can correctly classify all of the given data tuples. Intuitively, however, we expect the hyper plane with the larger margin to be more accurate at classifying future data tuples than the hyper plane with the smaller margin.

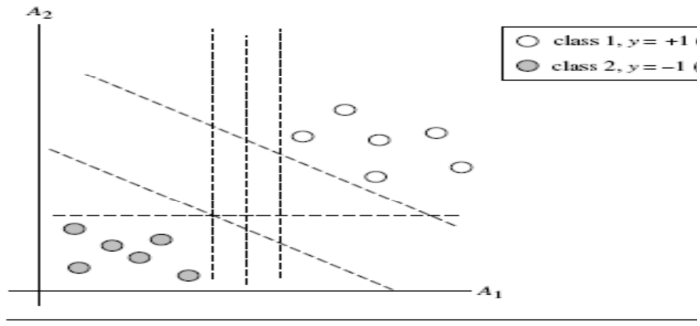


Figure 3. 7 linearly separable 2-D training data

This is why (during the learning or training phase), the SVM searches for the hyper plane with the largest margin, that is, the maximum marginal hyper plane (MMH). The associated margin gives the largest separation between classes. Getting to an informal definition of margin, we can say that the shortest distance from a hyper plane to one side of its margin is equal to the shortest distance from the hyper plane to the other side of its margin, where the "sides" of the margin are parallel to the hyper plane. When dealing with the MMH, this distance is, in fact, the shortest distance from the MMH to the closest training tuple of either class.

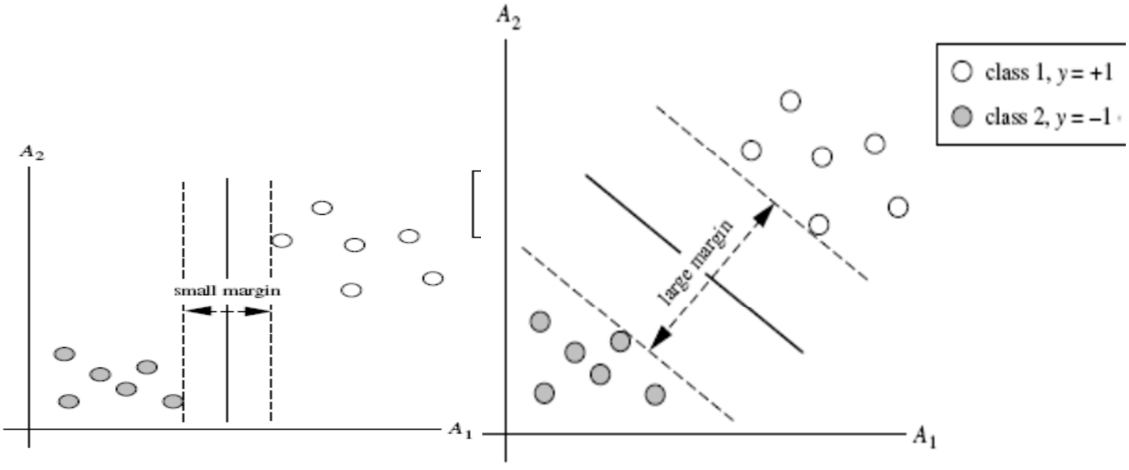


Figure 3.7 (a) (b)

2.5.1.4. 2.Constructing of Optimal Hyper plane Algorithms

Han and Kamber (2006) also reported that how the optimal separating line be drown based on mathematical perspectives. As a result a separating hyper plane can be written as

W*X+b = 0…….. ……..….(3.1)

Where W is a weight vector, namely = {w1, w2… wn}, n is the number of attributes and b is a scalar, often referred to as a bias. Besides, let's consider two input attributes, A1 and A2, as in Figure 3.7(b). The training tuples in 2-D, e.g., X = (x1, x2), where x1 and x2 are the values of attributes A1 and A2, respectively, for X. If we think of b as an additional weight, w0, we can rewrite the above separating hyper plane as

w0+w1x1+w2x2 = 0……..… (3.2)

Figure 3.7 tries to show two possible separating hyper planes and their associated margins. Which one is better? The one with the larger margin should have greater generalization accuracy. Thus, any point that lies above the separating hyper plane satisfies

w0+w1x1+w2x2 > 0…….…. (3.3)

Similarly, any point that lies below the separating hyper plane satisfies

w0+w1x1+w2x2 < 0……...(3.4)

The weights can be adjusted so that the hyper planes defining the "sides" of the margin can be written as H1 : w0+w1x1+w2x2 >= 1 for yi = +1,…….(3.5)

and H2 : w0+w1x1+w2x2 <= - 1 for yi = -1……(3.6)

That is, any tuple that falls on or above H1 belongs to class +1, and any tuple that falls on or below H2 belongs to class -1. Combining the two inequalities of Equations (3.5) and (3.6), we get yi(w0+w1x1+w2x2) >= 1, ∀i……. (3.7)

Any training tuples that fall on hyper planes H1 or H2 (i.e., the "sides" defining the margin) satisfy Equation (3.7) and are called support vectors. That is, they are equally close to the (separating)

MMH. In figure 3.7(b) the support vectors are shown encircled with a thicker border. Essentially, the support vectors are the most difficult tuples to classify and give the most information regarding classification.As Han and kamber (2006) stated linear SVMs we studied would not be able to find a feasible solution here for linearly inseparable data. But there is an approach to extend the steps described for linear SVMs with little modification to create nonlinear SVMs for the classification of linearly inseparable data. Such SVMs are capable of finding nonlinear decision boundaries (i.e., nonlinear hypersurfaces) in input space.

As a result the extended nonlinear SVM methods follows two major steps described below. In the first step, we transform the original input data into a higher dimensional space using a nonlinear mapping. Several common nonlinear mappings can be used in this step. Once the data have been transformed into the new higher space, the second step searches for a linear separating hyper plane in the new space. We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The maximal marginal hyper plane found in the new space corresponds to a nonlinear separating hyper surface in the original space.

3.5.1.4.3. Support Vector machine Algorithms

The Support Vector Machine (SVM) algorithm is probably the most widely used kernel learning algorithm. It achieves relatively robust pattern recognition performance using well established concepts in optimization theory (Cortes and Vapnik, 1995).

There are many kernel functions in SVM, but selecting good kernel function is a research issue. Basically, the followings are some popular kernel functions: linear kernel, RBF kernel, polynomial kernel and hyperbolic tangent kernel. Among this popular kernel functions, the researcher choose the linear kernel function because: the problems are linearly separable, so we don't need other types of kernel function. Besides the class size is two and dataset is easily manageable.

Hence, this research deals with linearly separable data, the researcher used Sequential minimal optimization (SMO) algorithms. Because it is available in WEKA and to implement this we use polynomial kernel function with degree 1 was used .The detail be discussed in experimentation part.

3.5.1.4.4. Sequential minimal optimization

Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines. It was invented by Platt in 1998 at Microsoft Research (Platt, 1998). SMO is widely used for training support vector machines and is implemented by the popular libsvm tool (Chang, C. et al, 2010).

SMO is an iterative algorithm for solving the optimization problem by breaking the problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers α_1 and α_2, the constraints are reduced to:

$$0 \leq \alpha_1, \alpha_2 \leq C \text{…….(3.8)}$$

$$y_1\alpha_1 + y_2\alpha_2 = k \text{ ……...(3.9)}$$

where C is an SVM hyper parameter and K(xi, xj) is the kernel function, both supplied by the user; and the variables are Lagrange multipliers. and this reduced problem can be solved analytically. The algorithm proceeds as follows:

1. Find a Lagrange multiplier that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.

2. Pick a second multiplier and optimize the pair.

3. Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence.

In general, According to Kotsiantis (2007) point out that the following are the general pseudo code for SVM algorithms are depicted as follows.

1. Introduce positive Lagrange multipliers, one for each of the inequality constraints (1). This gives Lagrangian:

$$\boldsymbol{l_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (x_i \cdot w - b) + \sum_{i=1}^{N} \alpha_i} \ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

2. Minimize Lp with respect to w, b. This is a convex quadratic programming problem.

3. In the solution, those points for which α i are called "support vectors" > 0

Even though the maximum margin allows the SVM to select among multiple candidate hyper planes, for many datasets, the SVM may not be able to find any separating hyper plane at all because the data contains misclassified instances. The problem can be addressed by using a soft margin that accepts some misclassifications of the training instances (Veropoulos et al. 1999). This can be done by introducing positive slack variable Si, i =1,..., N in the constraints, which then become:

w.xi – b≥ $+1 - \epsilon \; for \;\; yi = +1$…….(3.11)

w.xi - b≤ $-1 + \varepsilon \;\; for \;\; yi = -1$……(3.12)

$\varepsilon \geq 0$, Thus, for an error to occur the corresponding ε_i must exceed unity, so $\sum_{\mu i} \varepsilon_i$ is an upper bound on the number of training errors. In this case the Lagrangian is:

$$lp \equiv \frac{1}{2}\|w\|^2 + C \sum_i \varepsilon_i - \sum_i \alpha_i \left\{ y_{i\,(x_i .\; w - b) - 1 + \varepsilon_i} \right\} - \sum_i \mu i \varepsilon_i \quad \ldots\ldots\ldots \; (3.13)$$

where the μi are the Lagrange multipliers introduced to enforce positivity of the ε_i .Nevertheless, most real-world problems involve non- separable data for which no hyper plane exists that successfully separates the positive from negative instances in the training set. One solution to the inseparability problem is to map the data onto a higher- dimensional space and define a separating hyper plane there. This higher-dimensional space is called the transformed feature space, as opposed to the input space occupied by the training instances.

Above all this research also investigated using SMO algorithms based on the recommendation of the previous researcher about the domain area and to boost and test the performance of the classification technique of SVM against the status of HIV for a given client.

3.5. 2.Clustering

Clustering is another task of data mining in which groups of instances or objects that are more similar belong to the same clusters whereas dissimilar to different clusters (Han and Kamber, 2001). The only difference between classification and clustering is in the case of clustering unlike classification there are no predefined classes. As there are no predefined classes and examples in clustering, records are grouped together on the basis of similarity among the instances (Beaza-yates and Ribeiro-Neto, 1999).

3.5.3. Association rule

Discovers frequently occurring item sets in a database and presents the patterns as rules. This technique has been applied in problems like network intrusion detection to derive association rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks (Lee et al., 1999).Association learning is a data mining scheme in which any affinity grouping between features is sought, not just one that predict a particular class value (Agarwal and Srikant, 1994).

Association is suitable if the problem is to extract any structure that exists in the data at hand. The aim of association is to examine which instances are most likely to be grouped together. Market basket analysis is a typical practical application of association, which is concerned with what items are consumed with a particular item by individual buyer. Association rules can be developed in order to determine arrangements of items on store shelves in a given supermarket so that items often bought together be found arranged closer in location (Berry and Linoff, 1997).

3.6. Successful Data mining

The two basic things to be successful in data mining is to come up with the a precise formulation of the problem you want to solve and using the right data. The more the model builder can play with the data, build models, evaluate results, and work with the data some more in the given time, the better the resulting model become. Consequently, the degree to which a data

mining tool supports this interactive data exploration is more important than algorithms it uses (Two crows Corporation, 2005).

3.7. Potential Application of data mining

The application of data mining spans various industries. Telecommunications and insurance industries make use of data mining techniques to detect fraudulent activities. In medicine, data mining is used to predict the effectiveness of surgical procedures and medical tests.

Companies in the financial sector use data mining to determine market and industry characteristics as well as to predict individual company and stock performance, Predict which customers buy new policies ;Identify behavior patterns of risky customers (Two Crows Corporation, 1999).

3.8. Review of Related Work

There are a number of researches done to apply data mining techniques in health care domain in general and HIV/AIDS control and prevention program by extracting interesting pattern and rule in supporting the decision maker in particular.

3.8.1. Application of data mining in Health care Industry

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation (Palaniappan and Awang,2008).As a result, Medical institution may use the technologies developed in the new interdisciplinary field of knowledge discovery in databases (KDD) (Larvac, 1998), and particularly data mining (Lloyd-iams, 2002). Today, numerous health care organizations are using data mining tools and techniques in order to raise the quality and efficiency of health-related products and services. A number of articles published in the health care literature shown the practical application of data mining techniques in the analysis of health care related records.

As part of this effort the health care organizations are implementing data mining technologies to help minimize costs of clinical test, provision of quality service (diagnosing patients correctly and administering treatments that are effective) at affordable cost and improve the efficiency of patient

care. Besides, data mining can be used to help predict future patient behavior and to improve treatment programs.

By identifying high-risk patients, health professional can better manage the care of patients so they will not be problems of tomorrow (Rogers and Joyner, 2002).Recent trends in medical decision making show awareness of the need to introduce formal reasoning, as well as intelligent data analysis techniques in the extraction of knowledge, regularities, trends and representative cases from patient data stored in medical records (Larvac,1998).

According to Larvac, machine-learning methods have been applied to a variety of medical domains in order to improve medical decision-making. Razak and baker (2006) have also investigated a study focusing on mining association rules from asthma patients profile dataset. The purpose of the study was to identify attributes that affect asthma patients. The dataset of the asthma patients records were about 16,384 records and 118 variables in various formats. These attributes are grouped into demographic attributes and asthma related attributes.

They employ data preparation and association rule mining phases of data mining method. Understanding the nature of the dataset, identifying data types and formats, identify incomplete data, analyzing data distribution, and discretizing data are the important stages needed in order to systematically preprocess the data. As result of data preprocessing and cleaning only 31 attributes are left to generate association rules. The association rules mining phase uses Apriori Algorithm. To implement the association rule mining technique, testing and training dataset by determining threshold value is the main work of the researcher.

Palaniappan and Awang (2008) have investigated that how to predict the probability of patients getting a heart disease given patient records. In order to achieve this data mining goals, they used CRISP-DM methodology and data mining Extension (DMX) query language for graphical user interface design. They developed an intelligent system that predicts a given patient heart disease status. To develop a model the study used decision tree, naïve- byes and Neural network data mining techniques. The study used a total of 909 records with 15 medical attributes are used. The records are split equally into training dataset (455 records) and testing dataset (454 records). For

the sake of consistency, only categorical medical attributes are transformed to categorical data. The attribute diagnosis" is identified as the predictable attribute.

The prototype intelligent heart disease prediction system (IHDPS) developed based on the combined output of decision trees, naïve Bayes and neural network data mining techniques. The results of the study are encouraging and realize the objectives of the defined mining goals. The system extracts hidden knowledge from a historical heart disease database. The models are trained and validated against a test dataset. All model to have extracted patterns in response to the predictable state. The registered performance of naive byes is better to predict patients with heart disease and followed by neural network and then decision trees.

3.8.2. Application of data mining in HIV/AIDS

Various local and international scholars have attempted to study the application of data mining on HIV related records or data. The following works are reviewed by the researcher to help clarify the significance of the research problem and to show the gap in the current local and international studies.

Vararuk et al. (2008), for example, have investigated the application of data mining techniques on HIV/AIDS records to determine the pattern that exists in patient. The researchers patterns obtained in their study can be used for better management of the disease and appropriate use of resources in Thailand. The study took 250,000 records from HIV/AIDS patients. IBM's intelligent miner is used for clustering and association rule mining is applied to identify symptoms that may follow a set of existing ones.The study result of Vararuk et al.(2008) have showed clustering of group of patients with common characteristics . Association rules generate that were not expected in the data and are different from traditional reporting mechanisms utilized by medical practitioners. It also allowed the identification of symptoms that co-exist together. A good indicator of this research was that identification of symptoms that are forerunner of other symptoms can allow the targeting of the former so that the latter symptom can be avoided.

Similarly, Madigan, et al. (2008) have examined the association between the health workforce, particularly the nursing workforce, and the achievement of the millennium development goal, taking into account other factors known to influence health status, such as socioeconomic

indicators. Specifically, the main goal of the study was to investigate the association between health workforces with HIV prevalence rate at country level (worldwide). The dataset obtained from WHO and NGO sources are merged into one file at the country level for analysis. CART Methods over clustering data mining technique is used for prediction of the levels of HIV/AIDS prevalence rates among the 194 countries for which data are collected. The predictive model developed indicates that increased physician and nurse density (number of physicians or nurses per population) was associated with lower adult HIV/AIDS prevalence rate, even when controlling for socioeconomic indicators. Finally the performance of the model revealed that the main determinate factors in understanding HIV//AIDS prevalence rates are physician density followed by female literacy rates and nursing density in the country.

Elias (2011) conducted his research to determine the status of clients being HIV positive and negative for data collected from Addis Ababa Administration HCT users. To predict client's status the researcher used decision tree, J48 algorithms and Association rule of Apriori algorithms. To develop a model he followed CRISP –MD methodology. For preparation of dataset Weka 3.7.5 , and Microsoft Excel was used. The research output indicated that 81.8% performance was registered to associate the attribute and HIV status. Finally, the researcher recommends the implementation of the same research problem with other data mining techniques by increasing the number of dataset, number of variable to register better performance and develop knowledge based system based on the extracted rule to scale up HIV testing in Addis Ababa.

Teklu (2010) has attempted to investigate the application of data mining techniques on anti-retroviral treatment (ART) service with the purpose of identifying the determinant factors affecting the termination /continuation of the service. This study applied classification and association rules using, j48 and apriori algorithms respectively. The study was conducted using 18,740 ART patients' dataset. The methodology employed to perform the research work is CRISP_DM. Finally, the investigator proved that the applicability of data mining on ART by identifying those factors causing the continuation of the service. Birru (2009), has also investigated the applicability of data mining on VCT taking the case of CDC. He used 56,486 dataset from 2002 to 2008. The dataset contains unbalance HIV positive and negative clients' data and after the dataset is balanced only

14,793 records are considered for his experiment. To develop a model he employed CRISP-DM methodology and used clustering and classification data mining techniques.

Among Clustering techniques, k-means and Expectation maximization (EM) algorithms are used to define group of similar VCT client and to see how these grouped affect the classification outcome. From the two clustering algorithms EM is selected and the cluster indices created using this algorithm is then used as class for the classification purpose. In implementing the classification he has used J48, random tree and multi-layer perception to predict level of risks of clients as high risky or low risk based on the clustered indices. The performance of the model indicate that decision tree have shown better performance and appropriate to the domain. This is due to the fact that decision tree algorithm has a simple feature which can be easily understanding by non-technical staff. Above all the researcher recommended further research using large dataset and other data mining technique to boost the performance of the model.

Correspondingly, Abraham (2005) has conducted data mining research that examine the application of data mining technology to identify determinant factor of HIV infection and to find their association taking the case of CDC. The researcher used 18,646 records extracted from database of CDC. This dataset consists of 82 attribute, among them only 19 attributes were selected for training and testing the model. To develop a model the researcher used decision tree and association rule discovery. Knowledge STUDIO and Weka 3.7.5 data mining tools are used to implement the model and to extract rules to identify risk factor. Abraham has tried different association rule mining experiments by reducing attributes which the algorithm can do. He has attempted to find risk factors using only general association rule which it can bring any attribute in the consequent of the rule which may not be useful to identify most risk factors.

In general, the aforementioned review highlight that much has not been done and left out recommendation to try by other researcher in way to improve the performance of the model and get the actual problem solved. In this regards, this research has been attempted to determine variables that has significance impact on HIV infection. Consequently, the researchers have used 20 attribute, three different classification model, and different target population (no previous

studies has investigated the behavior of University community towards HIV with this technology) to attain the research objectives.

Hence, to the knowledge of the researcher, this research problem has not been attempted before and novel because more variables are used for training and testing the model; some of the features of the variables or attributes are different (such as location, residence and category) and lastly some of the classification model such as Support vector machines are not been used yet to investigate the determinate risk factor of HIV infection.

Taking this fact in mind, this research has initiated and conducted ,and would be expected to contribute a lot to managers and decision makers in HIV/AISD prevention and control practicing in VCT center.

CHAPTER FOUR

DATA UNDERSTANDGING AND PREPROCESSING

4.1. Overview of African AIDS Initiatives International

AAII (African AIDS Initiative International) is a Non profitable and Non-Governmental Organization (NGO) and is registered both in the State of Massachusetts, U.S.A and Ethiopia working with the motto of "Fighting for Life". (LLGP, 2007).

AAII envisions HIV/AIDS free, healthy and productive Africans. The mission of the organization is preventing the incidence of HIV infection and mitigating the impacts of AIDS among Africans. The goal of the organization is developing and promoting a culture of excellence for HIV/AIDS prevention, service delivery, networking and research.

To achieve its goal, the organization aims at providing HIV/AIDS education for positive behavior change, creating access to Voluntary HIV/AIDS Counseling and Testing (VCT) services, provide care, support and treatment for people living with HIV/AIDS (PLWHA) and Orphans, spearheading and facilitating collaborative partnership among the national and international actors working on HIV/AIDS in Africa. AAII opened its Regional Office for Africa in Addis Ababa in April 2001. It is strongly believed that the AAII's Ethiopia operations will serve as a model to replicate its activities in other African countries in collaboration with African Union (AU), Harvard AIDS Institute and World Health Organization (WHO) (LLGP, 2007).

As part of its effort to contain the spread and mitigate the impact of HIV/AIDS at national level, AAII is involved in various HIV/AIDS prevention and control activities. In this regard AAII is involved in the provision of VCT services, support to students' and workers' associations, girls' and reproductive health (RH) clubs, provision of quality HIV/AIDS/RH services through establishing resource centers, care and support to PLWHAs in the various faculties and schools of Addis Ababa University. Apart from its activities in the academic institutions, AAII is involved in providing care and support to PLWHAs and AIDS orphans in Mojo, Debre Birhan, Debre Markos

and Addis Ababa. Besides, it provides support to commercial sex workers in selected areas of Addis Ababa, particularly (Arada Sub City).The intention of this research is to evaluate the practice of the university community towards HIV infection insides and outsides university(area where university community exposed to their personal entertainment) campus .

As a result, this research mainly focuses on VCT service provided in Addis Ababa university community and centers that resides in Arada Sub city; particularly centers that are found in the surrounding of Addis Ababa University. The VCT service provided in different faculty mainly in Arat kilo and Sedist Kilo campus and off campus center which is located in Arada sub city around Minillek Hospital. The process of VCT service is discussed in chapter three, section 3.4.1 in detail. The counsellor provides both pre-testing and post testing counselling. According to the counsellor of AAII they spent much time in pretesting phase in discussing issue related to HIV/AIDS. Once the client decided to get test the counselling process takes place in the following ways.

They initially fill client information in standard paper format manually see **Appendix 1.** The format has three parts, the first part will be filled by receptionist, the second part called "pre-testing" which is filled by the counsellor and the last part called "post testing" filled by counsellor too. After pre-testing phase, the counselor takes the client to the laboratory for test if and only if the client decided to take the HIV test. The information gathered from clients is collected from all VCT centers to be encoded centrally in computerized system. The software is called Epi Info system which is capable of handling medical related VCT data. The documentation, data entry and analysis are performed in the data management section. This section is located in off campus site and has one IT (Information Technology) professional to encode data and handle manipulation of the system.

The proximity and being staff of the University, the researcher's get access to and collect the required data for this study encouraged the researcher to conduct this research at AAII.

4.2. Business Understanding

The main goal of AAII is the realization of an HIV/AIDS free community at all operational level sites where men and women in general and youth in particular are empowered to use their potential to control the spread and mitigate the impact of HIV/AIDS. AAII achieves its objective

empowering youth through enhancing the knowledge, attitude/behavior and life skills of the student to detect (high) risk activates that leads to acquisition of and transmission of the HIV pandemic. Besides, expansion of VCT service in campus as well as off campus site is the main organizational goal. According to the counsellor and data manager of the AAII, in campus site various means are used to teach the university community to mitigate the spread of HIV/AIDS in campus.

These are conducting consecutive training for trainer to the selected student member, distribution of condom, and leaflet through student representative across dormitory, student lounge and within their office. Besides, pear education is the best mechanism that the organization is currently using to create awareness about the incurable disease. This mechanism basically works with one student for a minimum of five students (1:5) is assigned and the assigned student is responsible for introducing about the disease for his/her specific group. She/he can be consider as a "preacher" and as a result s/he gets a pocket money for accomplishing the assigned task.

On the other hand, the organization also provides voluntary counselling and testing service in off campus site to address the service to university surrounding and commercial sex worker community.

4.2.1. Classification of AAII intervention area

AAII just like other international as well as local organization has its own five years strategic plan to accomplish its mission. It is a licensed and registered international organization. Its license is renewed yearly based by Ministry of Health. It has established a framework to be followed under five major intervention areas. These are:

1. **Prevention education:** Providing HIV/AIDS education to create awareness and bring behavioral change. Providing financial assistance for student anti-AIDS club and worker and female students club.

2. **Condom promotion and distribution**

3. **Voluntary counselling and testing**: to create the opportunity and access to quality VCT services. In this regards AAII establish four VCT sites: Sidist kilo, Science, Mobile VCT

center and Off campus(to provide access to those who are not comfortable to use in campus service)

4. **Care and support**: It also provides financial assistance especially for sex worker and economically unaffordable client. Because of lack of sufficient fund the AAII set up an agreement with university surrounding government hospital such as Black lion, yekatit 12 and Menilik for providing ART(Antiroviral Treatment) service.

5. **Capacity building and Networking and research.** Establish database and network facilities for disseminating information. Facilitating a networking and partnership among the VCT community.

To accomplish its task the main source of fund AAII is from USAIDS, UNAIDS, Care International, HAPCO (HIV/AIDS Prevention and Control Office) and Federal HAPCO**.** Even though, such fund are earned through competition of project proposal the AAII has the following three major problems.

a) Renewal of licenses in annual base is too bureaucratic and time taking. As a result of this the organization missed project funder.

b) Lack of transportation facility such as car.

c) Lack of sufficient budget to expand the scope of VCT service. As discussed above, the organization provides only VCT service, after the client result is positive, it doesn't provides financial support for that client.

In a normal circumstance AAII provides voluntary counselling and testing service to the community in campus and off campus site. The organization only told the laboratory result of clients after being tested in either site. Beyond that there is no way of identifying variable that cause HIV infection rather keeps the client information in computerized system. Basically AAII analyze reports or the activities of the VCT services using traditional methods of data analysis such as MS Excel (mean, mode, variance etc.). However, such methods of data analysis has limited capacity to discover new and unanticipated patterns and relationship that are hidden in

conventional databases (Plate et al.,1997). Besides, those identifying patterns of attribute or risk factors of HIV infection are difficult because the dataset contains or involves too many attribute or parameter. So studying the patterns of the attribute that have relationship among each other rather than listing are more crucial and support decision making.

Consequently, applying DM technology can handle and generate such pattern of variables that cause HIV infection and used the output by concerned party to facilitate decision making process. Hence, this is main factor that the researcher initiated to conduct this research.

4.3. Understanding of Data

According to Cios et al. (2007) model, the next phase after understanding the business and problem domain has been data understanding. Hence, a prerequisite for undertaking DM research was dealing with data itself. For this reason, Berry and Linoff (2004) described that the identified good source of data to attain the DM goals has been the corporate data warehouse. Hence the data stored in corporate database (AAII VCT database in this context) maintain consistency, in their attribute value, data types and common format.

4.3.1. Data Source and Data Collection

There are two key tasks in data mining. The first one is coming up with precise formulations of the problem you are trying to solve. The second task is using the right data (Two Crows Corporation, 2006). Of the two main tasks analyzing and understanding the content and structure of the collected data is one of the most important tasks that need attention in data mining process.

The source of data for this research has been collected from AAII VCT center database. The data base was manipulated by using Epi info software system. It was standard medical software that suite for handling medial related data. It has a user interface that allows the data encoder to enter clients' information. There were 18,800 clients records in the database but the researcher used only 15,396 for the reason that some records are noise ,incomplete, and out of the scope of the study. All the records are taken from the database starting from 2005 to 2011 G.C. The summary of the data sources is illustrated below in Table 4.1

Table 4. 1 Description of data source and Number of records

Source of data center	Data coverage	Number of records	Number of attribute	Size of the data	Data Type
AAII VCT data	2005-2011 G.C	15,398	63	4.2MB	Nominal

4.3.2. Description and Quality of Data

The number of attributes of collected data from the VCT center is about 63. However, with the discussion of the domain expert, only relevant attributes is selected to perform data mining tasks. The selected attributes are described in Appendix 2. According to Han and Kamber (2006) data quality can be measured in terms of accuracy, completeness, consistency, timeliness, believability and interpretability. In this regard, VCT data suffer from incompleteness, missing value, invalid value and inconsistency problems. Since data quality can also be affected by the structure of the data being analyzed, the researcher give due emphasis to this aspect of the data too. In the next section only the selected VCT data for experimentation is summarized statistically relative to the whole data.

4.3.3. Descriptive Statistical Summary of Selected attributes

This initial dataset has been described and visualized using Microsoft Excels to examine the properties of the dataset relative to the whole records. Simple statistical analysis has been performed to verify the quality of the dataset such as missing values, error values and to obtain high level information regarding the data mining questions. Hence, the selected attributes used for model building are statistically described in details below. This is helpful for understanding of the dataset for experimentation.

4.3.3.1. Attribute Selection

In this study VCT data has been used, however, the original records were full of problems and errors; they were not used directly for data mining purpose. The process of filling client's information manually consists of three phases. In the **first phase**, there are 19 attributes to be filled

at reception when clients came to the sites for VCT services. These are "**Country", "Region","Woreda", "Site code", "Site type", "Organization type", "Resident", category", "year ", "Campus", "Client code", "Return visit", "New code", "Age", "Sex", "Counselor code", "Partner code", "Couple code", and "Visit date".**

Out of this only **"Year"," Campus", "Residence", "Age" and "Sex"** attributes were selected, because all of them directly related to the problem to be solved to attain the research objectives. However, Country and Regions have been removed because contains single value attribute and has no role for prediction to the subject under study.

Moreover, "Site code", "Site type", and "Organization type "attributes have not been used to identify the type of the centers but have not relevant for this particular problem. Similarly "Client code"," Counselor code and "Partner code" were not used since they are unique values and not relevant in showing patterns. "Return visit" and "New code" attributes were used to check whether the client has been returned to the center or new, but these attribute was redundant instead" previous test attribute" was used.

Furthermore, pre counselling phase consists of 25 attributes (see Appendix 1), out of which 15 of them has been selected. Other attributes such as how you hear of about VCT, from where the client is referred, and date previously tested were not considered as relevant for the purpose of this research. All attributes in the post-test counseling session that are obtained after knowing the HIV status were not considered since they did not help in predicting the HIV infection risk factor of the clients. Thus, 20 attributes are selected for experiments.

The selected final attributes were statistically described below and see the details in Appendix 2.

Residence attributes. It stands for the place where client came from. The attribute has 12 values and they were nominal.

The frequency of the attribute with its various values looks like as depicted in Table 4. 2. A client who came from Addis takes 74.41 % of the whole dataset. The missing taken about 1.64 %, however; this value was filled by the most frequented value which was Addis.

Table 4. 2 Summary of Residence's Attribute

Residence : Nominal		
Distinct values	**Frequency**	**Percentage (%)**
10-Addis	11458	74.41
11-Dire	59	0.38
1-Region 1	720	4.68
2-Region 2	19	0.12
3-Region 3	1178	7.65
4-Region 4	1167	7.58
5-Region 5	14	0.09
6-Region 6	31	0.20
7-SNNP	422	2.74
8-Region 8	16	0.10
9-Region 9	31	0.20
Others	31	0.20
Missing Value	252	1.64
Total	15396	100

Category attribute: The frequency of the attributes and the corresponding possible nominal value has been shown in Table 4.3. As shown in table, the modal value for category was regular student. It has eleven values. The missing value of the attribute was 9867 (64 %). It takes more than half of the total records percentage. According to Han and kamber (2006) applying the most frequent value (here regular student) is used to fill the missing value.

However, when we observe the actual records their residence value is Addis (urban). So discussing with the domain expert the missing value is filled with community. This is because the number of the off campus client is greater than in campus client.

Table 4. 3 Statistical summary of category attribute

Category : Nominal		
Distinct value	**No of clients**	**Percentage (%)**
Regular student	5854	38.02
Extension student	359	2.33
Summer	403	2.62
Staff	967	6.28
Community	845	5.49
Student partner	306	1.99
other	1133	7.36
Missing value	9867	64.08
Total	15396	100.00

Year Attribute: It is a nominal attribute having ten distinct values including 1st year to 8th year and the other which are not applicable is another value. By examining the current university curriculum, the researcher reframes the distinct value as follows in Table 4.4.

As shown in the following Table 4.4 , the common university community who taken VCT service was third year but the community by far greater because their status belong to non - student. It is about 57% of the total population. Basically the initial data indicate that both staff and community were contained a missing and very few with not applicable value.

Based on the reality and convenient for experimentation result the researcher filled the missing value not with modal value rather with not applicable.

Table 4. 4 Statistical description of the year attributes

Year :Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Not applicable	402	2.61
1st year	1326	8.61
2nd year	1399	9.09
3rd year	1725	11.20
4th year	1595	10.36
5th year	123	0.80
6th year	33	0.21
Missing	8795	57.12
Total	15396	100.00

Missing value of the records in the above table registered 57.12 % of the total selected dataset. But, the actual records of the missing value was belongs to non-student and found unfilled (missing) similarly to Table 4.3, missing value is filled with not applicable value instead of 3rd year student which is relatively the most frequent value.

Sex Attribute: It contains two valid values that are either male or female. The detail is presented in table 4.5. As shown in table 4.5, male clients (63.39%) are more than female clients (36.52%).This implies more number of male than female clients was visiting the centers.

Session type Attribute: This attribute indicates whether the clients come to visit the center individually or in couple. It has four possible value namely individual, group, couple and other. Table 4.6 presents the distribution of clients by session type.

Table 4. 5 Statistical summary of Sex attribute

Sex :Nominal		
Distinct value	Frequency	Percentage (%)
Male	9761	63.39
Female	5623	36.52
Missing Value	14	0.09
Total	15396	100.00

Table 4. 6 Statistical summary of session type Attribute

Session Type: Nominal		
Distinct Value	**Frequency**	**Percentage (%)**
Individual	13410	87.09
Couple	1891	12.28
Group	9	0.06
Other	9	0.06
Missing value	79	0.51
Total	15396	100.00

Marital status Attribute: The Attribute of marital status contains seven distinct values such as married, never married, separated, divorced, widowed, living together and other. The frequency of the attribute with possible nominal value will be described in Table 4.7

As shown in the table below, modal value for marital status was never married. The missing values of the attribute were 90 (.58%) which is very little so it better to fill with the modal value.

Table 4. 7 Statistical Summary of Marital status Attribute

Marital status: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Married	1514	9.83
Never married	13187	85.64
Separated	129	0.84
Divorced	290	1.88
Widowed	121	0.79
Live together	7	0.05
Other	60	0.39
Missing value	90	0.58
Total	15396	100.00

Education Attribute: This attribute reveals the level of education of clients. It has nine distinct values (illiterate, PHD, Able to read, Primary, Secondary and other mentioned in following Table 4.8).

Employee Attribute: This attribute has a nominal value and refers to the client's status who employed or not. The distribution of the client about their employment situation has been described in Table 4.8. It has two distinct value such as Yes and No.

Table 4. 8 Statistical Summary of Employee Attribute

Employee: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
0-No	9092	59.05
1-YES	6143	39.89
Missing value	163	1.06
Total	15396	100.00

Table 4. 9 Statistical Summary of Education Attribute

Education: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Illiterate	249	1.62
PHD	3	1.94
Able to read	245	1.59
Primary	1686	10.95
Secondary	3969	25.7
Diploma	1640	10.65
B.Sc.	452	2.94
Undergraduate St.	6366	41.34
Postgraduate St.	382	2.48
BA degree	158	1.03
Other	94	0.61
M.A degree	2	0.01
Missing value	152	0.99
Total	15396	100.00

As shown in table 4.9, clients whom educational status was undergraduate student and visiting the centers most frequently accounts 41.34% of total dataset. It followed by clients who has secondary education level (25.78%) and primary educational levels (10.95%) respectively .Some missing (0.99%) values were observed in AAII VCT data

In the above table we can visualize that the percentage of client being employed was less than form not employed. The percentage of not employed versus employed is 59.05 and 39.89 % respectively. The remaining percentage is missing value which is 163(1.06%).

Primary Reasons Attribute: This attribute describes the reason why the client is visiting the center. There are around 21 distinct values for primary reason for using the VCT service. The frequency of the distinct value with their respective percentage value will be shown in the Table 4.10 below.

Table 4. 10 Statistical Summary of Primary Reasons Attribute

Primary Reasons: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
2nd Test(win)	833	5.41
Confirm positive result	70	0.45
Get results previous test	2	0.01
Need Counseling	453	2.94
Test before pregnant	9	0.06
Pregnant must know	28	0.18
Plan for future	9322	60.54
Death/illness of partner	48	0.31
Occupational exposure	54	0.35
Other blood/Fluid exp.	294	1.91
Client Risky/Had risk	1462	9.49
Sexual assault	14	0.09
To know status	70	0.45
Partner Risky/Had risk	180	1.17
Not trust partner	397	2.58
Ill/Symptoms	174	1.13
Premarital	409	2.66
Marital reunion	31	0.20
Family planning	3	0.02
Visa application	142	0.92
Other	1120	7.27
Referred	7	0.05
Missing value	**245**	**1.59**
Total	**15396**	**100.00**

As shown in table 4.10, most of the clients who visiting the centers for planning their future (which accounts for 60.54%) followed by when their client was at risky condition (9.49%). Moreover, 245(**1.59**%) missing values were identified from the data.

Previous Tested Attribute: This attribute refers to whether clients have been previously tested or not. The clients visiting the centers may have come previously and have different previous HIV status or they may be coming for the first time. The summary of the previous test status of the clients is presented in table 4.11**.**

Table 4. 11 Statistical Summary of Previous tested Attribute

Previous Test result: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
No	5968	38.76
Yes, HIV+	204	1.32
Yes, HIV -	8998	58.44
Yes, inconclusive	13	0.08
Result not given	2	0.01
Didn't take results	5	0.03
Other	13	0.08
Missing values	195	1.27
Total	15396	100.00

Table 4.11 indicate that even though, there have been various reason for getting HIV test, getting negative result encouraged clients to come again in the VCT Center (Which accounts 58.44%) and followed by clients not previously tested (it accounts 38.76 %).

Ever Had Sex Attribute: This attribute shows whether the client has ever had sex since s/he was born or not**.** The summary of the data are presented in table 4.12

Table 4. 12 Statistical Summary of Ever Had Sex Attribute

Ever Had Sex Type: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
No	5525	35.88
YES	9734	63.22
Missing Values	139	0.90
Total	15396	100.00

As shown in table 4.12, most clients (63.22%) had experienced sex when they come to visit the centers. There is 0.90% missing values.

Suspected Exposure Time: This attribute is the length of time between the date of the client's most recent possible exposure to HIV and the date of the VCT visit. Table 4.13 shows the detailed distribution of clients by exposure time.

Table 4. 13 Statistical Summary of Suspected Exposure Attribute

Suspected exposer Time: Nominal		
Distinct value	**Frequency**	**Percentage (%)**
No exposure	3708	24.08
<1 month	1298	8.43
1 to 3 months	1209	7.85
4 to 6 months	2035	13.22
Over 6 months	2998	19.47
Don't know	2883	18.72
N/A	1093	7.10
Missing values	174	1.13
Total	**15396**	**100.00**

As shown in table 4.13, the suspected exposure time accounts of 24.08% clients who has no exposure and followed by Over 6 months (which accounts 19.47%) when they visit the center. Moreover, 1.13% missing values is detected from the data.

Condom Use Attribute: It refers to the condom usage manners of clients during sexual intercourse for the last 3 months. The frequency distribution of clients are illustrated in table 4.14

Table 4. 14 Statistical Summary of Condom Use Attribute

Condom use: Nominal		
Distinct value	**Frequency**	**Percentage**
Never	5236	34.00
Always	2134	13.86
Sometimes	770	5.00
N/A	7081	45.99
Missing values	177	1.15
Total	**15396**	**100.00**

As illustrated in the above table 45.99% of the clients are observed as not applicable for use of condom and those who accounts about 34% are never use the condom. Additionally, 177(1.15%) of the client use condom behavior for the last six month is treated as missing value or unfilled.

Used condom Last Sex Time attribute: This nominal attribute is talk about the client's condom use behavior during last sexual intercourse with sexual partner. It has four possible distinct value namely: - yes, no, does not remember and N/A (not applicable). The frequency distribution is depicted be in Table 4.15

According to Table 4.15 indicated that 46.7 and 31.79 % of the clients have not used condom during last time they had sex, however 19.74 % have used condom. Minimum number of missing value (1.09%) is identified from the records.

Table 4. 15 Statistical Summary of Used Condom last Attribute

Condom use Last Time had sex: Nominal		
Distinct value	**Frequency**	**Percentage**
No	7191	46.70
Yes	3040	19.74
Doesn't remember	104	0.68
N/A	4895	31.79
Missing value	168	1.09
Total	15396	100.00

History of Sexual Transmitted Infection (HSTIs**) attributes:** This attribute deals about client's personal history of has been make a diagnosis or practice of STIs. It has same distinct value of as "condom use last time had sex" attribute. The detail is presented in table 4.16.

Table 4. 16 Statistical Summary of History of Sexual Transmitted infection Attributes

H STI: Nominal		
Distinct value	**Frequency**	**Percentage**
No	10263	66.65
Yes	150	0.97
Don't know	13	0.08
N/A	4769	30.97
Missing value	203	1.32
Total	15396	100.00

In the above table we can observe that 66.65% of the clients have not known their sexual transmitted infection history. The number of missing value is very few and accounts about 1.32%.

Steady Attribute: A numeric attribute that indicate the number of permanent sexual partner that clients have ever contacted with. It described in terms of mode, mean, outlier and missing value. Statistical distribution of the number of sexual partner for a given client be illustrated below in Table 4.17.

Table 4. 17 statistical summary of steady attribute

Distinct Value	Frequency	Percentage(%)	Remarks
Invalid value	**6261**	**40.66**	**1-,-01,-,_,**and *
0	1990	12.92	
1	6969	45.26	
2	18	0.12	
3	4	0.03	
4	8	0.05	
5	1	0.01	
6	1	0.01	
7	**1**	**0.01**	
10	**3**	**0.02**	
14	**5**	**0.03**	
Missing value	**137**	**0.89**	
Total	**15396**	**100.0**	

Table 4.17 shows that the number of client who has permanent partner account of about 45.26% and followed by those who have no partner or zero value. The number of error or outlier and missing value detected from the data is 40.66 and 89% respectively. On the other hand almost half of the records were filled by invalid value.

Age attribute: The age attribute contains a numerical valid value in the range from birth to 99 years old client. Clients at various ages are coming to the centers for HIV screening. The details of the age of the clients are presented in table 4.18. As the researcher observed, there is no age missing value from the selected dataset.

More over the average age who involved in VCT service is 25. This indicates that they were the youngest and productive force of the community. Hence, the dataset includes both age groups

starting form 0 year to older age because the organization provides service for both university community and people who are found in surrounding of Addis Ababa University. The maximum age of client who took the VCT service is 85.

Table 4. 18 Statistical summary of age attribute

Age : numeric		
S,N	**Age distribution**	**Age number**
1	Minimum	0
2	Maximum	85
3	Mode	20
4	Average	25
Total number of Age in the dataset		**15396**

Location /Campus attribute: This nominal attribute records the client place of work or their affiliate faculty in which they registered. It is classified under nomenclature of the university faculty and arada sub city. The frequency of the clients, where they are living ,and where they takes VCT service depicted below in Table 4.19.

As shown below in table 4.19 majority of the client who took VCT service have not mentioned their campus or place where they currently settled. Which account about 55.26% considered as a missing value. But about 28.07 % of the client was living in sedist kilo. The some value of the attribute duplicated with different spelling of same name. For instance, ‘Art’ and ‘art school’, registered with different name but same value. It was corrected and categorized as one distinct value.

Table 4. 19 Statistical Summary of place of location/ campus of the client

Campus/Location : Nominal		
Distinct value	**Frequency**	**Percentage (%)**
Art	1	0.01
Pharmacy	1	0.01
Sedist kilo	4322	28.07
Science	2134	13.86
N.Technology	129	0.84
S.Technoligy	32	0.21
Commerce	21	0.14
Veterinary	5	0.03
B.Lion	4	0.03
ART School	3	0.02
Yarede	2	0.01
N/A	51	0.33
Other	83	0.54
B.Lion	1	0.01
Missing Value	8509	55.26
Total	**15396**	**100.00**

Similarly, as Table 4.3 and 4.4, here is a missing value which was handled using other than most frequented value. Because the campus only hold the value of the student rather than community or client. When we observed the actual data set the missing value belongs to "Addis". So the researcher with much discussion with domain expertise decided to fill using NA value.

Causal Attribute: This numeric attribute shows the number of all those sex partners the client has faced only once or rarely for the last 3 months. Table 4.20 summarizes the frequency distribution of the attribute causal. As shown below in table 4.20 the number of erroneous treated as causal value is 69.21%. The missing value is 1.07%. The average number that a client faced a sexual partner is less than one which is account of .95 % and it is followed by usual one partner contact.

Table 4. 20 Statistical summary of Causal Attributes

Causal: Numeric			
Number of Causal	**frequency**	**Percentage (%)**	**Remark**
0	1979	12.85	Minimum contact------0 Maximum contact------20 Average----------0.95 Missing value----------1.07% Error----------69.21%
1	2005	13.02	
2	302	1.96	
3	194	1.26	
4	43	0.28	
5	20	0.13	
6	11	0.07	
7	6	0.04	
8	1	0.01	
10	6	0.04	
11	2	0.01	
12	2	0.01	
15	3	0.02	
18	1	0.01	
20	1	0.01	
Missing value	165	1.07	
Invalid value	10657	69.21	--,-,-*,*,-+,0-,,1-,4--,4- are outlier
Total	**15396**	**100.00**	

This takes 13.02 % of the total client. Besides, the outlier that is wrongly entered in the record should be removed and corrected manually based on the proximity of the value by the researcher judgment.

HIV Test Result Attribute: This nominal attribute determine the client HIV status by laboratory test of his/her blood. The laboratory test will be investigated and determine by lab technician as HIV negative and positive. The overall investigation of the AAII data set regarding HIV status is illustrated in Table 4.21.

As depicted in the table 4.21 about 96.10 % of the client has be tested and grouped as HIV negative, 3.07 % of the clients are HIV positive. The number of missing value and error detected from the data are .79% and 0 .01% respectively.

Table 4. 21 Statistical summary of HIV Test Result Attributes

HIV test: Nominal		
Distinct value	**Frequency**	**Percentage**
Negative	14798	96.10
HIV+	473	3.07
NA	4	0.03
Invalid	1	0.01
Missing value	122	0.79
Total	**15396**	**100.00**

4.4. Data Preparation and Preprocessing

These steps consist of the core of data mining and take much of the research time and effort of the entire knowledge discovery process but critically, the most compulsory activity in data mining process (Han and Kamber, 2001).

Data preprocessing involves all the action taken before the actual data analysis process start. It can be defined as a transformation that transforms the raw real world data to a set of new data (Famili and Turney, 1997). It is unrealistic to expect that data will be perfect after it is extracted. Since good models usually need good data, a thorough cleansing of the data is an important step to improve the quality of data mining methods.

As Berry & Linoff (2004) have pointed out, the process of preparing data is like the process of refining oil. Oil starts out as a thick tarry substance mixed with impurities. It is only by going through various stages of refinement that the raw material becomes usable. Just as the most powerful engines cannot use crude oil as a fuel, the most powerful algorithms (the engines of data mining) are unlikely to find interesting patterns in unprepared data. The correctness and consistency of the data also matter the performance of data mining research. So as to boost the performance of the algorithms, the researcher tried to assess such problems which require due emphasis by applying different preprocessing techniques. Such methods are data cleaning, data

reduction, and data integration and transformation. Each of these methods will be discussed in details as follows.

4.4.1 Data Cleaning

Data Cleaning is a process which fills in missing values, removes noise and corrects data inconsistency. Usually, real world database contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2001).Consequently, data cleaning has become a must in order to improve the quality of data so as to improve the performance of the accuracy and efficiency of the data mining techniques.

This technique involves removing the records that had incomplete, noise (invalid) data and filling missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire dataset.

As a result , the researcher make use of MS-Excel 2010 built-in functions like search and replace, filtering, and auto fill mechanisms ,and WEKA to identify and fill missing value.

4.4.1.1. Missing Value Handling

Missing values refer to the values for one or more attributes in a data that do not exist. In real world application data are rarely complete. It can also be a particularly pernicious problem. Especially when the dataset is small or the number of missing fields is large, not all records with a missing field can be deleted from the sample.

Moreover, the fact that a value is missing may be significant itself. A widely applied approach is used to calculate a substitute value for missing fields, for example, the median or mean of a variable (Famili and Turney, 1997). Accordingly, the researcher has been analyzed the VCT dataset and identified missing values and take measure to solve the problem as follows. As we depicted in section 4.3.3, missing values are clearly identified and calculated the percentage of the value against total records. As a result, the entire selected attribute illustrated less than one percentage (1.00%) except campus ,

year and category attribute which of them accounts of 55.26%, 57.12%, and 64.08% respectively. To handle the problem of missing values for nominal data type attributes, replacing with modal value is recommended in Two Crows Corporation (1999).For numeric data type attribute the missing values were recommended to be replaced by the mean of that value a whole

Therefore, based on the above principles the researcher handles the missing value and WEKA preprocessing replace missing value techniques also used. WEKA fills using the most frequent (modal) value methods which is same as the above principle. Additionally, manually tracing and fixing the missed value is other technique used by researcher. Summary of handled missing value is illustrated in Table 4.22

As shown below in table 4.22 out of the selected 20 attributes 19'n of them have registered with a missing values. Accordingly, the researcher reacts to take appropriate action to clean the data.

4.4.1.2. Handling outlier value

The data stored in a database may reflect outlier – noise, exceptional case, or incomplete data object and random error in a measure of variable. These incorrect attribute values may be due to data entry problems, faulty data collection, inconsistency in naming convention or technology limitation (Han and Kamber, 2001). The authors have also explained four basic methods for handling of noise data. These are binning method, clustering, regression and combined computer and human inspection.

In this research, the researcher has identified and detected noise or outlier value from the VCT data. With the help of domain experts the identified outlier was corrected manually. As mentioned in the previous section, we tried to describe the outlier particularly in table 4.18 and 4.21 of steady and causal attribute. Accordingly, a combined effort of the researcher and domain expert were taken to identify and correct the problem of the inconsistence, incompleteness, noise or outlier.

Table 4. 22 Summary of handling missing value

Number	Attribute name and their data type	% of missing value	Replaced with	Reason/Technique applied
1	Residence :Nominal	1.64	Addis	Most frequent value
2	Category: Nominal	64.08	Community	Reason discussed under table 4.4
3	Year : Nominal	57.12	Not applicable	Reason discussed under table 4.5
4	Sex:Nominal	0.09	male	Most frequent value
5	Mar status: Nominal	0.58	Never Married	Most frequent value
6	Education :Nominal	0.99	Undergraduate	Most frequent value
7	Employment:Nominal	1.06	no	Most frequent value
8	Reas Hear:Nominal	1.59	plan	Most frequent value
9	PrevTest:Nominal	1.27	no	Most frequent value
10	EverHSex:Nominal	0.9	yes	Most frequent value
11	SusexpTime:Nominal	1.13	No experience	Most frequent value
12	Conduse:Nominal	1.15	Not Applicable	Most frequent value
13	LastCondUse:Nominal	1.09	no	Most frequent value
14	HSTI:Nominal	1.32	no	Most frequent value
15	Steady: Numric	.89	1	Most frequent value
16	Causal:Numeric	1.07	1	Most frequent value
17	SessType: Nominal	.51	Individual	Most frequent value
18	Campus:Nominal	55.26	Not applicable	Reason discussed under table 4.20
19	HIV test:Nominal	0.79	negative	Most frequent value

In table 4.20 we identified 10657("-") incomplete values the researcher used bin mean smoothing to replace the missed data as recommended by Han and Kamber (2006).

With this regards the mean of the total records is 0.95 and rounded to 1. so the replaced value will be “1”. By the same token, some value of the records exhibit inconsistency and different from the existing one which is called noise data or outlier. For instance, “1” and “01”,2 and “02”, “many” and ” too many” and other reveal inconsistence and ”-+”,” _*”,”3--,4--”,-1,-0 and much other outlier are observed from this particular attribute. In similar fashion the researcher remove 0 from prefix and keep the remaining data as a value of the record and remove many and too many value and replace with maximum value i.e. “20”. On the other hand, based on the recommendation of Han and Kamber (2006), we removed the outlier “-+”,” _*” and replace with the most frequent value (mode) which is “1” and remove suffix (- -) from “3- -“and “4- - “and keep their value as it is.

Moreover, in table 4.17, we can see same problem aforementioned above and same technique is applied to clean the data before implementation using WEKA.

4.5. Data transformation and Reduction

As Han & Kamber (2006) have stated, data mining often requires data integration or the merging of data from multiple data sources. The dataset which were used for the subsequent model building are prepared and derived from one source, AAII VCT clients records, hence no need of merging of the database.

In data transformation, the collected data are transformed or consolidated into forms appropriate for mining. The process of data transformation might include smoothing (using bin means to replace data errors), normalization, where the attribute data are scaled so as to fall within a small specified range (scaling a data inside fixed range), attribute construction, where new attributes are constructed and added from the given set of attribute to help the mining process (Chackrabarti, et al., 2009). The data needed to be reduced in order to make the analysis process manageable and cost- effective. Chackrabarti et al.(2009) and Han and kamber (2006) discussed in their books that data reduction techniques include data discretization is one of data transformation methods which is used to reduce the number of values for a given continuous

attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values, data cube aggregation, dimensional reduction (irrelevant or redundant attributes are removed), and data compression data is encoded to reduce the size, numerously reduction (models or samples are used instead of the actual data).

In this research some attributes were discretized to reduce the unlike values of the attribute to obtain knowledge (pattern) and to make the dataset suitable for data mining tools. As a result, the following section depicted the details of them.

4.5.1 Discretization and concept hierarchy generation

Han and Kamber (2006) reported that raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Discretizing the residence attribute value: this attribute consists of the nine Ethiopian regional state (Afar, Mekele, Harari , SNNP, Gambela , Benshangul, Oromiya, Amhara and Tigray) and the two City administration namely Addis and Dere. But some value is repeated with different name with same distinct value. Here it replaced with addis and the remaining distinct value is transformed starting from R1 through R9. For addis, Dere and other option as R14, R10 and 99 are used respectively.

Discretization summary of year attribute: Based on the current curriculum of tertiary education, the researcher tried to associate and classify the current academic year of user of VCT. For this reason the year category of the client will be range and transformed as Y1 to Y6 and “NY” instead of ’not applicable’ and ‘others’ are substituted.

Discretizing the category Attribute value: So as to understand with a high level of concept and easy to interpret the output of data mining, and gain advantage of it, the researcher discretize the category attribute value as follows in Table 4.23.

To decrease the size of the attribute (same value of the attribute with different values and simplicity of manipulation) the researcher tried to collocated and transform in to new value. For instance ‘staff family’ and ‘community’ almost consist of same behavioral value. So it will be redundant and should be removed.

Table 4. 23 Discretization summary of Category attribute

Attribute Name	**Old value**	**Transformed value**	Remark
category	Regular student, post graduate and student	Regular student	For the sake of easy understanding and interpretation, the most related values are grouped together.
	Extension student	Extension student	
	Summer student	Summer student	
	Staff	staff	
	Community , Staff family	Community	
	Partner	Partner	
	Other, Visitors and not applicable	Other	

Discretization summary of Education attribute: The different distinct value of educational attribute has been transformed in to high level of concept. The details has been shown in Table 4.25

Moreover, the researcher also tried to transform the distinct value of the campus or location. The old and transformed values were shown below in table 4.24

Table 4. 24 Campus attributes values discretization

Attribute Name	**Old Attribute value**	**Transformed Value**	**Remark**
Campus	Sidist Kilo	Sidist Kilo (SS)	Social science and FBE members are included
	Technology N. & S.	Technology N. & S. (Technology)	The number of Tech s.is very few , if u merged , no problem created
	Commerce	Commerce (Com)	
	Medicine	Medicine(Med)	B.lion and B.Lion , both of them belongs to same family
	Science	Science(SC)	Art and Art schools holds same data and under the constitute of science
	NA, other	Not Applicable	Hence, NA is a high level and inclusive to incorporate other.

Table 4. 25 Education Attribute values Discretization

Attribute Name	**Old Values**	**Transformed value**
Education	Illiterate	Illiterate
	Able to read	Able to read
	Primary	Primary
	secondary	secondary
	undergraduate	undergraduate
	postgraduate	postgraduate
	other, TVET, diploma and other certificate is included	other

As shown in table 4.25 the value of the attribute transformed and grouped in to high conceptual level. For the sake of analyzing the data easily using WEKA, it is also converted the transformed value in to nominal. The new grouped distinct value is"other" which transformed from "diploma","TVET" and the old name"other" educational level. Therefore, the final valid distinct values will become seven valid distinct values.

Discretizing the Reason Hear Attribute value: The primary reason the clients are visiting the centers has 22 distinct attribute values. These attribute values are grouped in to five higher knowledge level values. These are Plan, Risk, Causal, Know status and other. The details high level grouping will be illustrated below in Table 4.26

As shown in table 4.26, the researcher transformed the details old value into concrete five new values to improve the performance of the models.

Table 4. 26 Discretization of ReasHere Attribute Values

Attribute Name	**Old values**	**Transformed value**	Remark
Reason Here	"Client Risky" ,"Partner Risky", "Not trust partner" , "Other blood/fluid exposure", "Sexual assault", "Preliminary ART"	Risk	
	"Premarital" , "Marital reunion", "Family planning", "Visa applicant", "Need counseling", "Test before pregnant", "Pregnant must know", "Plan for future"	Plan	
	"TB". "death Illness of partner", "Ill/Symptoms" , "Occupational exposure",	Causal	Clients exposed to HIV causal
	"Referred" , "2nd Test (windows) " , "Confirm positive result", "Get results previous test", "Status know"	Status know	
	Other	Other	

However, from the context of HIV/AIDS, clustering this possible distinct value in to high knowledge level and removing unnecessary data help both the researcher and the machine to easily manage and generate appropriate pattern.

Accordingly the researcher by discussing the matter with domain expert grouped the value in to three distinct values. These are “No partner”, “One partner” and “Multiple partners” for both attributes. Hence, the actual values 0 and 1 are replaced by “No partner” and “One partner” respectively, whereas the value more than 1 is replaced by “Multiple partners”.

Discretizing the Values of Age Attribute: Replacing numerous values of a continuous attribute by a smaller number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining result. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts(such as numerical values for the attribute age) with high level concepts for instance Child hood, adolescent ,adult and old(Han and Kamber,2006). Accordingly, binning methods was used to scale the values of “Age” attribute. Such a meaningful discretized values used in this research are grouping the age value in to six distinct value including “0-15”, “16-20”, “21-25”, “26-35”, “36-45”, and “46+”. These groups shows pattern of “age” if any. Therefore, the actual values of the ages are manually replaced by the corresponding high level grouped values.

Table 4. 27 Discretization of age attributes value

S.N	Age attribute in range	Transformed value
1	“0-15”	A1
2	“16-20”	A2
3	“21-25”	A3
4	“26-35”	A4
5	“36-45”	A5
6	“46+”	A6

As shown in the above table range of age attribute value reduced in to six distinct high conceptual levels.

Once the data sets are pre-processed and cleaned, the next steps is develop a data mining model using WEKA 3.7.5 software which is handled in the next chapter.

CHAPTER FIVE

EXPERMINTETION AND ANALYSIS OF RESULT

In this chapter, the researcher describes the techniques that have been used in developing a model to predict HIV infection risk factor. This research incorporated the typical stages that characterize a data mining process. This study has been organized according to hybrid processing model, which is described and discussed in section 1.4.1 and 2.4.4 and shown in figure 3.3 of chapter one and two respectively. Here the researcher discuss the experimentation process by relating the steps followed ,the choice made , the task accomplished , the result obtained, evaluation of the model and results , and present it in a way that the organization can easily understand and use it.

5.1. Model Building

Modeling is one of the major tasks which is undertaken under the phase of data mining in hybrid methodology. In this phase several data mining techniques are applied and their parameters are adjusted to optimal values. Typically, different techniques can be employed for similar data mining problems. Some of the tasks include:- selecting the modeling technique, experimental setup or design, building a model and evaluating the model.

5.1.1. Selecting Modeling Technique

Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification techniques has been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, PART, J48 and SMO are used for experimentation of this study.

The researcher selected the above algorithms, easy of understanding and interpretation of the result of the model.

A PART algorithm is common types of rule induction technique which generate a model as a set of rules. In the meantime, the J48 algorithms of decision tree generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class output (Witten and Frank, 2005). Besides, SMO algorithm is available algorithm for Support vector machine implementation in WEKA to classify high dimensional data.

J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA (Two Crows Corporation,1999).It is WEKAs implementation of C4.5 top-down decision tree learners proposed by Quinlan(1986). This research used C4.5 algorithms to predict HIV infection risk factor. This algorithm is implemented by modifying parameters such as confidence factor, pruning and unpruning, changing the generalized binary split decision classification and other option available in Table 5.1. Therefore, it is very crucial to understand the available options to implement the algorithms, as it can make a significance difference in the quality of the result. In many cases, the default setting will prove adequate, but to compare results /models and attain the research objectives other opitions are considered (Witten and Frank, 2005, Han and Kamber, 2006).

Table 5. 1 Parameters for building J48 trees.

Name	Possible values	Default value	Description
M- Number of Instance	1, 2…	2	Minimum number of instances in leaves (higher values result in smaller trees).
U_ un pruned tree	yes/no	no	Use un pruned tree (the default value 'no' means that the tree is pruned).
C. confidence factor	10^{-7}- -0.5	0.25	Confidence factor used in post pruning (smaller values incur more pruning).
S-sub tree raising	yes/no	yes	Whether to consider the sub tree raising operation in post pruning.
B-use binary splits	Yes/no	no	Whether to use binary splits on nominal attributes when building the tree.

As indicated in above Table 5.1 the commonly used parameter of J48 algorithms was illustrated with various options related to tree pruning. Pruning produces fewer, more easily interpreted result. In other word, pruning can be used as a tool to correct for potential over-fitting. The algorithms recursively classifies until each leaf is pure, meaning until dataset has been categorized as close to their respective classes possible. Hence, the process ensures the maximum accuracy on the training data, but creates excessive rules.

In general, Witten and Frank (2005) explained the various option of tree pruning in J48 algorithm as follows. Basically, J48 uses two pruning methods; these are subtree replacement and subtree raising. The former method used J48 in which the nodes in a decision tree may be replaced with a leaf to reduce the number of tests along a certain path. This process starts from the leaves of the tree, and works backwards toward the root. The later methods used where a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtree raising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtree raising can be somewhat computationally complex.

Reduced-error pruning option: Error rates are used to make actual decisions about which parts of the tree to replace. There are multiple ways to do so. The simplest method is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over fitting. This approach is known as reduced-error pruning. This method reduces the overall amount of data available for training the model.

Confidence factor option: Confidence factor is the approach that seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires

a confidence threshold, which by default is set to 25 percent. This option is important for determining how specific or general the model should be. If the value of the confidence factor is lowered on the selected model, the training data is expected to fit in closely to the data we would like to test. As a result of this a more pruned or more generalized tree will be produced.

Minimum number of instances per leaf option: this is the lowest number of instances that can constitute a leaf. The higher the number the more general the tree will be. Lowering the number will produce more specific trees, as the leaves become more granular.

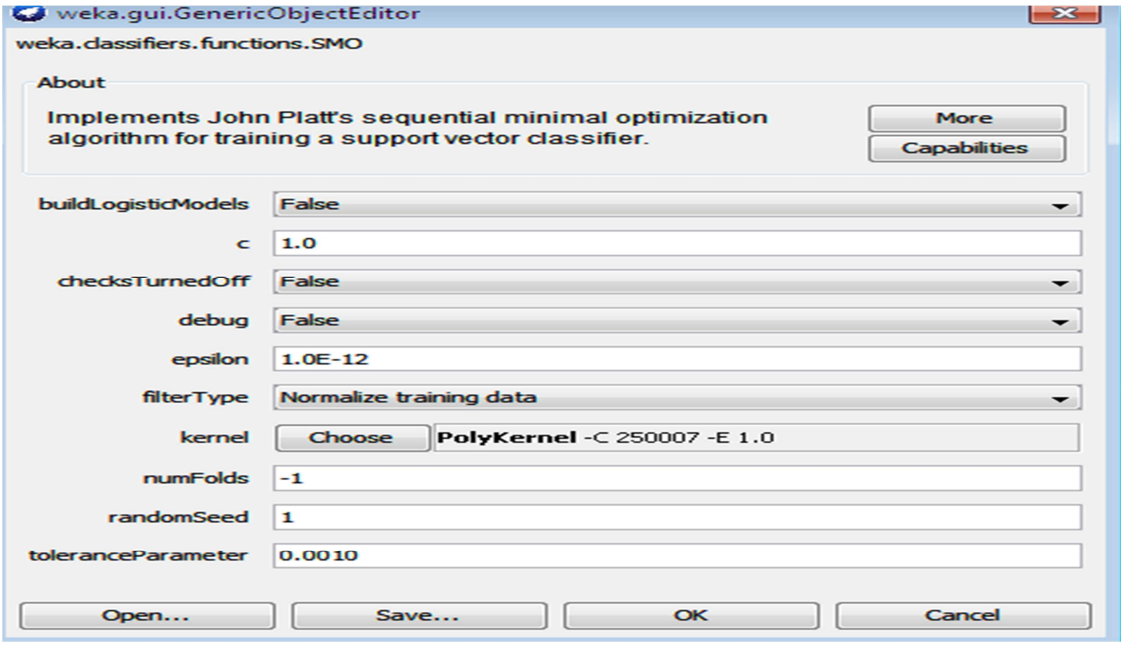
Binary split option: The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees.

Laplace smoothing option: This option is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities.

If the data mining researcher decides to employ tree pruning, it is advisable to consider the above options. Depending on how the training and test data have been defined, the performance of an unpruned tree may superficially appear better than a pruned one. This can be a result of overfitting. Hence, it is important to repeatedly experiment with models by intelligently adjusting these parameters to obtain the best set of options. According to Witten and Frank (2005), five basic and familiar parameter that usually used for data mining research is depicted in Table 5.1. PART algorithms are also separate-and-conquer rule learner proposed by (Witten and Frank, 2005). The algorithm generates sets of rules called ’decision lists ’which are ordered set of rules.

PART builds a partial C4.5 decision tree in each iteration and converts the “best” leaf into a rule. The five basic parameter mentioned for J48 algorithms decision tree are also available and applicable for PART. Since SVM popularity in 1998, it controls complexity and overfitting issues, so it works well on a wide range of practical problems (Vapnik ,1998). Due to this reason and previous studies research directions, the researcher selected to implement SVM in this study. To investigate the applicability for these research purpose SMO algorithms has used. SMO is nothing but implementation of SVM in WEKA. Before it gets to implement the software automatically convert all nominal attributes into sets of binary attributes. Then the WEKA takes the dataset as input.

Figure 5. 1 The snap shot of SMO algorithms



As shown in the above figure, the possible parameters used to implement the SVM are depicted. Besides, under kernel parameters there is also other possible parameter that is applicable for SMO to extend. The details description of the common basic parameters used in SMO are as depicted in Table 5.2. (Platt,1986).

5.1.2. Experimental Setup

In any data mining process before building a model, we need to generate a procedure or mechanism to test the model's quality and validity. For instance, in supervised data mining tasks such as classification, it is common to use classification accuracy measure or error rates as quality measures for data mining models.

Table 5. 2 Parameter for building SVM models

Name	**Default value**	**Description**
BuildLogistic Models	False	if this is true, then the output is proper probabilities rather than confidence scores
C-Complexity parameter	1.0	limits the extent to which the function is allowed to overfit the data
filterType	Normalize training data	whether you normalize the attribute values
Exponent	1.0	Used to describe the degree of the polynomial kernel
numFolds	-1	cross validation for training logistic models
Kernel	Polynkernel-c250007-E 1.0	Use to select types of kernel depending for data set(being linear and non-linear)

Besides, other standard measure including precision, recall, sensitivity and specificity are available. Therefore, the test design specifies that the dataset should be separated into training and test set, and builds the model on the training set and estimate its quality on the separate test set. With this regards Two crows Corporation (2005) reported that the process of building predictive models requires a well-defined training and validation protocol in order to insure that most accurate and robust prediction.

In this research 15,396 datasets are used for training and testing. WEKA 3.7.5 software has used to set up and measure the quality, validity and test of the selected model. For purpose of this study k-fold (10-folds) cross validation and percentage split test options are used because of its relatively low bias and variations (Witten and Frank, 2005).

Accordingly, the datasets are randomly partitioned equally into ten parts. Hence, 90% of the dataset is for training and 10 % for testing for former and the dataset are partitioned in to percentages (70-30 splits option meaning 70% of the dataset for training and remaining for testing).

In addition to these a different 3000 testing dataset is used for testing the prediction performance of the classification model developed. These testing dataset is prepared by systematic random sampling technique from the original dataset.

Witten and Frank (2005) reported that 10-fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier. To build the model of this research 19 independent and 1 dependent variables or attributes are used .The details of the selected attribute is discussed in chapter four and see the result in Appendix 2.

5.1.3. Attribute ordering

Since attribute selection is important in decision tree models, the researcher tried to rank the attribute based on information gain. It was calculated based on entropy value of the attribute. As Witten and Frank (2005) explained information gain is calculated from sum of entropy for every attribute. The formula for calculating intermediate values is:

$$\text{Info (D)} = -\sum_{i=1}^{m} Pi \log_2 Pi \qquad (5.1.)$$

Where, P_i is the probability that an arbitrary tuples in sets of training data D belongs to certain class. Info (D) is also known as the entropy of *D*. After you calculate information gain for each attribute, select the one with the highest information gain as the root node, and continue the calculation recursively until the data is completely classified by J48 algorithms in this case.

For the purpose of this research, WEKA is used to compute the information gain and rank according to the importance of the attribute for the classifier. As a result of this, the following

figure 5.2 depict ranked order of attribute based on their relevance for the reason that such attributes are very important for later experimentations by excluding the least relevant attributes.

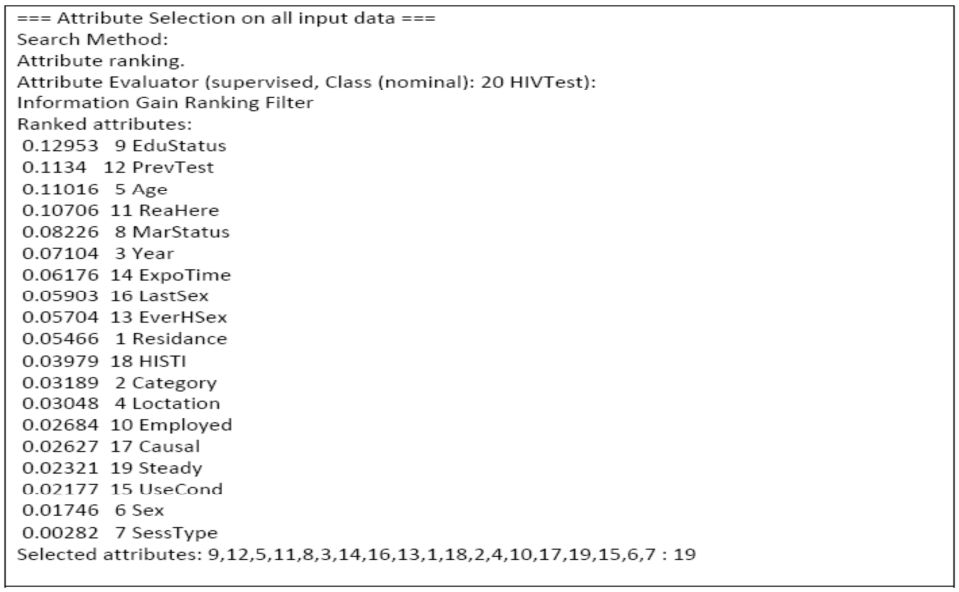


Figure 5. 2 Summary of ranked Attribute

As you can see from the result of attribute selection using entropy based information gain method of WEKA, out of 19, the top 15 the most relevant attributes are Edustatus, Previous test, Martialstat, Reasonhere, Ever had sex, Age, History of Sexual intercourse, Last sex without condom, year, residence, steady , location ,causal ,employees and exposure time.

5.1.4. Running Experiments

As it has shown in experimental setup of 5.1.2 for training and testing the classification model the researcher used two methods. **Methods One**: Percentage split method (holdout), where 70 % of the data used as training and the remaining 30 % testing. **Method Two**: K-fold cross validation methods, the data was divided into 10 folds, some fold is used as testing and the remaining folds are used as training.

Based on the above methods establishing scenario for model to be developed is very important to see the model result and analysis of each result; to compare the result of one model with the

previous one and finally help us to find out the outperforming model based on criteria of evaluation. Consequently for both of the methods following scenario has been done for each of three selected model with default parameter value of WEKA 3-7-5 software.

Scenario #1 : Pruning with all attributes

Scenario #2 : Unpruning with all attribute

Scenario #3 : Pruning with reduced attribute

Scenario #4 : Unpruning with reduced attribute

Once the modeling tool is chosen and performance evaluation criteria was established, then it followed by building model with a number of parameters that govern the model generation process. The choice of optimal parameters for the problem at hand is an iterative process, and it has to be properly explained and supported through results. So the next section present the resultant models that should be properly interpreted and their performance explained.

5.2 Model building using J48 decision tree

In Hybrid or Ciso.et al. methodology, data mining particularly model building is the crucial steps that should be deals with. Model building is an iterative process's. Therefore in this study different experiment is conducted altering the parameter with mentioned methods in section 5.1.4 using J48 algorithms, PART and SMO algorithms for building best predictive model. The parameters of each algorithm have been discussed in details in section 5.1.1.Using aforementioned methods; the experiment has been conducted with four (4) scenarios for each methods.

The researcher used 15,396 dataset with 20 attribute to investigate the model. Hence, in the real environment as everybody guess, the number of clients who have being infected with HIV is out weighted by those who have been free from the virus. This research dataset is not out of this estimation, so the actual dataset is 14,866 are HIV negative client and the remaining is positive.

However, conducting experiment with imbalanced dataset misleading classification performance and biased to the majority class (Chawla, 2003**)**. Additionally, it brings about poor decision to be taken as a result of the experiment. To coup up this and be fair the researcher used resampling technique to balance the imbalanced dataset in preprocessing phase with WEKA software. As described in chapter two of section 2.5.1.1 decision tree induction follows a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built (Han and Kamber, 2006). All of the 19 attributes, which are selected for model building, are fed as independent variables and the dependent variable of HIVTest is a target class. The algorithms builds starting at each node it will be sent either left or right according to some test. Eventually, it will reach a leaf node and be given the label associated with that leaf. In general, this research is more interested in generating rules that best predict the HIV infection and to come up with an understanding of the most important factors (variables) that can be a cause of HIV/AIDS. As mentioned in Table 5.1 J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Accordingly, based on two methods selected the following series of experiment has been conducted. Summary of experimentation with J48 algorithms result is depicted in Table 5.3

As mentioned in chapter one of methodology section 1.4.1.5, one of the compulsory step of hybrid methodology next to model building is evaluation of the model. Accordingly, the performance of the model has been evaluated based on the following criteria (see details in section1.4.1.5) including performance accuracy, confusion matrix value, and TP and FN rate, Number of leaves and size of the tree generated and ROC curves and execution time.

As shown below in Table 5.3 the experimentation has performed in two methods (in 70-30 percentage split and 10 -fold cross validation test option respectively.) The first experiment has be tested with four scenario mentioned above with 70- 30 percentage split criteria. When we compare the result of method I. Scenario # 2(unpruning with all attribute) is the best model based on correctly classified instance(out of 4619 , 4459 instances are correctly classified), performance register (95.7 % accuracy and TP rate of 96 % (classifying correctly HIV negative

client as HIV negative and HIV positive as positive). Besides, in terms of minimum number of leaves and size of tree (both indicate the complexity of generated tree) and time taken to build a models are Scenario # 4 (pruning with reduced attribute) and scenario # 3 (unpruing with reduced attribute) is best model respectively.

Table 5. 3 Experimentation result of J48 Algorithms with two methods

Model Characteristics	Experiment (Scenario)							
	1	2	**3**	4	1	**2**	3	4
Test option	**I**				**II**			
pruned	yes	no	**yes**	no	**Yes**	no	Yes	no
Attribute	All	All	**reduced**	reduced	**ALL**	ALL	reduced	reduced
Accuracy	**95.2**	**95.7**	**93.8**	**94.5**	**95.5**	**96.3**	**94.9**	**95.6**
Time taken	1.1	0.9	**0.88**	0.45	**1.5**	0.64	1.0	0.48
#of leaves	1707	2074	**1614**	2021	**1707**	2074	1614	2021
Size of trees	2106	2554	**1966**	2459	**2106**	2554	1966	2449
AV .TPR(%)	95	96	**94**	95	**95**	96	95	96
AV.FPR(%)	5	4	**6**	5	**5**	4	5	4
AV.PR	0.96	0.97	**0.95**	0.95	**0.97**	0.97	0.95	0.96
AV.RR	0.95	0.97	**0.95**	0.95	**0.97**	0.97	0.95	0.96
AV.ROC	0.97	0.98	**0.97**	0.98	**0.98**	0.98	0.98	0.98
CCI	4397	4459	**4331**	4364	**14706**	14833	14619	14722
ICI	222	160	**288**	255	**790**	563	777	674

Key: CCI: Correctly classified Instance, ICI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TPR: True Positive Rate. FPR: False Positives Rate, ROC: Relative Optical character curve, PR: precision rate, RR: Recall rate, I: 70-30 percentage split option (Holdout), II: 10- fold cross validation

The last compares in method one is made between Average ROC curve rate which registered a performance of 97 and 98 % in pruned and unpruned respectively irrespective of the number attribute. Above all from the method I, even though the complexity of the generated tree is high, correctly classifying instance and performance of the model is better .As a result Scenario # 2 (Building decision tree unpruning with all attribute) is selected.

Similarly, in the second experiment or method two (10-fold cross validation cases) four of the stated trial was conducted. When we compare the result of each scenario developed model, scenario # 2 (building decision tree unpruned with all attribute) is best based on accuracy which registered 96.3 % and correctly classified instance which is accounts 14,833 out of 15396 . Furthermore, based on criteria of minimum time taken for building model, scenario # 4 (building model with unprunned with reduced attribute) is best.

Analogously, the number of tree and size of tree for scenario # 3 is minimums so it is best because it reduced the complexity of the generated tree. The average ROC curve performance measure indicates same for all scenario and account of 98%. Finally when we compare the actual correctly classified instance based on their labeled class i.e. classifying HIV negative client as HIV negative for all scenario, irrespective of number of attribute unpruned register better than pruned. By the same token, from methods two, still scenario number two better than other model in terms of accuracy, correctly classified instance and ROC curve.

However, when we compare the size and leaves of tree of unpruned J48 model, the number is enormous and complex relative to pruned one. As a result of this the algorithms might not reach optimality and generate more generalized decision tree rules. Han and Kamber (2005) also reported this is because of overfitting problem. This is a fundamental problem in learning of algorithms. Besides, such situation has its own impact in classification performance particularly classifying unseen or new instance. Subsequently to solve the problem the researcher selected pruned scenario that perform better accuracy. Accordingly, scenario #1 (Building pruned

decision tree with all attribute) of 10- fold cross validation(method II) selected as the best J48 decision tree model. See the details confusion matrix in appendix 3.1

5.2.1. Comparison of J48 decision tree model

In general when we compare both methods in terms of classification model accuracy 10 fold cross validation (method II) is better than 70-30 percentage split .The accuracy ranges from 94.9 to 96.3 but 70-30 percentage split 93.8 to 95.5. Additionally, WEKA experimenter tool was used to compare of the J48 model developed with various scenarios. The result of the comparison is depicted in Figure 5.3

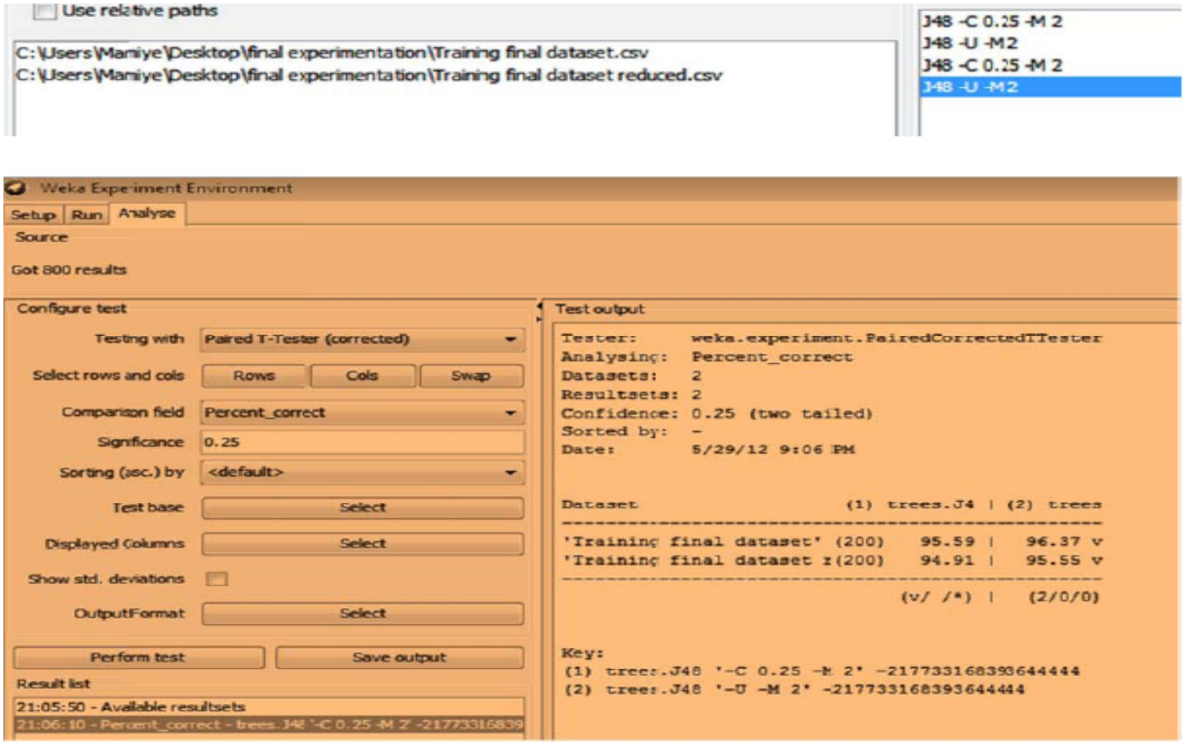


Figure 5. 3 Experimental dialog box with compered J48 algorithms

Consequently, as shown in the above figure and Table 5.3 experimentation, irrespective of reduced or all attribute the performance of unpruned J48 algorithms register better than pruned.

However, by the same token aforementioned, the researcher selected scenario # 1 (building decision tree pruned with all attribute) of 10 fold cross validation which registered **95.6**% accuracy as a final J48 decision tree model for training and testing the dataset and validated the performance of the model after getting reduced its complexity with increasing minimum instance parameters.

The detailed confusion matrix as a result of pruned J48 algorithm with all attribute has been discussed below in Table 5.4 and shown in Appendix 3.1

Table 5. 4 Summary of confusion matrix for pruned J48 decision tree with all attribute

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	7040	636	7676
	Positive	54	7666	7720
Total		7094	8302	15396

The base of this research is determining the risk factor of HIV infection and finding pattern that shows the status of the HIV. In this respect, the J48 algorithms classify how those instances are correctly and incorrectly classified in labeled class.

So as indicated in the above Table 5.4 based on the 10–fold cross validation test option the J48 learning algorithm scored an accuracy of 95.5%. This result shows that out of the total training datasets 14706 (95.5 %.) records are correctly classified, while only 690 (4.5%) of the records are incorrectly classified.

Besides, out of the total 7676 actually HIV negative clients, only 7040 (91.7%) clients are classified as HIV negatives and the rest are misclassified as HIV positive. And out of the total 7720 actually HIV positive clients, only 7666 (99.3%) clients are classified as HIV positive and

the rest are misclassified as HIV negative. This means the model has better performance in terms of correctly classifying HIV positives than HIV negatives.

According to Shermer (2002) report hypothesis testing, error can be categorized in to two based on the classification of being true as false and false as true. These errors are called Type I(false postive error) and Type II(False negative error) So in this research context of the experiment, classifying HIV negatives as HIV positive (Type II error) may lead the client anxiety and incurs cost of VCT testing. Whereas classifying HIV positives as HIV negative (Type I error) may have also its own risk such as spread of the disease. Hence, Type I error is more serious than Type II error in this case since it can affect more people.

Furthermore, evaluating the model based on sensitivity and specificity are very significance for decision making. For that reason .the result of the above confusion matrix indicate that the sensitivity of this test was (7666/7720) = 99.3 % and the specificity was (7040/7676) = 91.7%. The test indicates that the models appear to be pretty good. Because, based on the evaluation criteria, the classifier correctly classify clients as HIV negative who has actually diseases free with 91.7% accuracy and classify clients as HIV positive who actually has the deasese with 99.3%. However, the false positive and false negative rate which corresponds to classification error and register performance accuracy of 0.95 and 0.05 respectively.

As discussed in Table 5.3, the size of the tree and the number of leaves produced from method two of scenario # 1 model training was 2106 and 1707 respectively. This seems that it is difficult to navigate through all the nodes of the tree in order to derive out with valid sets of rule. Therefore, to make ease the process of generating rule sets or to make it more understandable, the researcher attempted to modify the default values of the parameters so as to minimize the size of the tree and number of leaves.

In this regards, the minNumObj (minimum number of instances in a leaf) parameter was tried with 10,20,30,50 100 and 200. However, the minNumObj set to 200 gives a better tree size and

minimal accuracy compared with the other trials. With this value of the minNumObj the process of classifying records proceeds until the number of records at each leaf reached 200. The result of the experiment depicted in Table 5.5 and the instant of the confusion matrix with minNumObj= 200 shown in appendix 3.2

Table 5. 5 Summary of pruned J48 algorithms with minNumObj = 200

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	6033	1643	7676
	Positive	1797	5923	7720
Total		7830	7566	15396

The experiment has shown an improvement in the number of leaves and tree size. The size of the tree is dropped from 2106 to 95 and the number of the leaves decreased from 1707 to 76. As we can see from Table 5.4, the resulting confusion matrix shows that the J48 decision tree algorithm scored 77.7 % accuracy. That means out of the total 15396 records 11956 (77.7 %) are correctly classified and the remaining are misclassified.

In other words, the confusion matrix of this experiment has shown that 6033 (78.6%) of the 7676 total HIV negative clients are correctly classified as disease free. But 1643 (22.4 %) of them are misclassified as HIV positive class. Besides the result shows that out of 7720 total HIV positive clients records 5923 (76.7 %) of the records are correctly classified while 1797 (23.3 %) of the records are misclassified as HIV negatives being HIV positives.When we compare this experiment result, classifying clients who has HIV positives as HIV positives outweigh from those classifying clients as HIV negative. Overall, nevertheless execution time reduce from 1.5 to 0.52 seconds ,and the tree size and the number of the leaves decreased significantly as shown above, the accuracy of the J48 decision tree algorithm in this experiment is poorer than the first

experiment with the default parameter value of the minNumObj. Besides, the accuracy of the model lowered from 95.5 to 77.7 % as minNumObj getting large which shown a considerable change. Hence the first selected algorithms is optimal and exhaustively see the whole thing irrespective of is efficiency. And can conclude that from this experiment minNumobj and accuracy of model performance have shown inversely related pattern. Additionally to keep the performance of the model the first experiment with the default minNumObj parameter value is better and taken as the best J48 decision tree model. The generated possible decision tree is shown in Appendix 4.

As described in section 5.1.2 randomly selected datasets are used to validate the performance of the model. Accordingly, once it has been developed the J48 model using the whole training dataset with 10 –fold cross validation test option and registered the possible capacity of the model that accurately classify new instance. However, validating the developed model with independent dataset will enhance believability and applicability of the model. To do so the researcher, using supplied test option re-evaluate the model by mentioned dataset. The result of the reevaluation of the model depicted as follows in Table 5.6 and sees the snapshot of the model in appendix 3.3

Table 5. 6 Result of re-evaluation of J48 decision tree with 3000 dataset

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	1409	103	1512
	Positive	73	1415	1488
Total		1482	1518	3000

The performance of the model with this testing dataset was 94.1%. This result shows that out of the 3000 testing datasets, the developed decision tree classification model predicted around 2789 (94.1%) records correctly.

In general, the researcher exhaustively discussed based on the result of the model with domain expert to determine variable that cause HIV infection. According to the experiment depicted in Table 5.4, the J48 model used only eleven attribute out of 19 attributes including PrevTest, EverHsex,Year,Age,Usecond,ReasHear, causal, Category, ExpoTime, Sex , and employed. So from this model, these are the determinate attribute of HIV infection. However, the researcher cross checked the validity based on expert opinion. According to Health expert, marstatus and edustatus attribute are also can be used as the cause of HIV infection.

Hence, the above mentioned attribute plus experts opinions (overall 13 attributes) are the most determinat factor of the HIV infection. Besides, the pattern /rules obtained by setting pruned, confidence factor 0.25 and test option 10 fold cross validation J48 decision tree with all attribute (in this case 19) model is useful for identifying HIV infection risk factor and pay attention on those important clients behavior by health professional and by the university management in particular.

5.3 Model building using PART Rule Induction Algorithms

The second data mining classification technique applied in this research was PART Rule induction algorithms. As mentioned in chapter two of literature review section 2.5.1.3,there are many rule induction algorithms but the researcher selected PART for the reason that PART has the ability and potential to produce accurate and easily interpretable patterns/ rules that helps to achieve the research objectives. PART is a separate-and-conquer rule learner like that of decision tree and proposed by Witten and Frank(2005).The algorithm is apply an iterative process and produces set of “decision lists” which is ordered set of rules. It works by generating a rule that covers a subset of the training examples and then removing all examples covered by the rules from the training set. This process is repeated until there are no examples covered left to

cover. The final rule set is the collection of the rules discovered at every iterations of the process. The rules are in standard form of IF-THEN rules.

To build the Rule induction model, WEKA software package and the same 15,396 VCT dataset was used as an input respectively. The experiment will be performed analogously as we did in former model. That means the experiment was divided in two methods as methods one: experiment with 70-30 percentage split criteria for four scenario and Method two as with 10-fold cross validation evaluate the four scenario again. The parameters are partially adjusted and default value was used with reduced and all attribute. Accordingly, the experiment of all scenarios with methods one will be illustrated below in Table 5.7

Table 5. 7 Experiment result of PART algorithms for two methods

Model Characteristics	Experiment (Scenario)							
	1	2	**3**	4	1	**2**	3	4
Test option	**I**				**II**			
pruned	yes	no	**yes**	no	Yes	**no**	Yes	no
Attribute	All	All	**reduced**	reduced	ALL	**ALL**	reduced	reduced
Accuracy	**95.7**	96.5	**94.7**	95.1	**96.7**	**97.3**	95.4	96.2
Time taken	**23.7**	37.3	**15.2**	24.7	24.0	**28.8**	15.5	24.5
No. of Rules	374	508	393	524	374	**508**	393	524
AV .TPR	0.96	0.97	**0.95**	0.95	0.97	**0.97**	0.95	0.96
AV.FPR	4	3	**5**	5	3	**3**	5	4
AV.PR	0.96	0.97	**0.95**	0.95	0.97	**0.97**	0.96	0.96
AV.RR	0.96	0.97	**0.95**	0.95	0.97	**0.97**	0.95	0.96
AV.ROC	0.97	0.98	**0.97**	0.97	0.98	**0.98**	0.98	0.98
CCI	4421	4459	**4373**	4392	14886	**14975**	14687	14814
ICI	198	160	**246**	227	510	**421**	709	582

Key: CCI: Correctly classified Instance, **ICI** (Incorrectly classified Instance), **Accuracy**: Registered performance of model, **AV**: Average, **TPR**: True Positive Rate. **FPR**: False Positives

Rate, **ROC**: Relative Optical character curve, PR: precision rate, **RR:** Recall rate, **I**: 70-30 percentage split option (Holdout), **II**: 10- fold cross validation

As shown in the above table, the registered performance is better that induction rule learner in case of unpruned instead of pruned one. But by same token aforementioned in case of J48 decision tree classifier, the researcher selected the one which register best with pruned case irrespective of the number of attribute and methods.

Consequently, among the four scenarios experimented with 70-30 percentage split (method I) with varying parameters, scenario #1 registered better performance of 95.7%. This shows that out of the training set of 4619 records 4421 (95.7%) of the records are correctly classified, while 198 (4.3%) of the records are misclassified.

		Predicted HIV test result		
		Negative	Positive	Total
HIVstatus (Actual)	Negative	2145	195	2340
	Positive	3	2276	2279
Total		2148	2471	4619

Table 5. 8 Confusion matrix of PART algorithm with 70-30 percentage-split

Furthermore, the resulting confusion matrix also shown in Table 5.8 that out of the total 2340 HIV negative clients 2145 (91.7%) of them are correctly classified in their respective class, while 195 (8.3%) of the records are incorrectly classified in the HIV positive classes or wrongly classified client as HIV positive without having really the diseases). In other hand out of the total number of HIV positive clients 2276(99.86%) of them are correctly classified as HIV infection as positive but 3 (0.14%) of the client records are misclassified.

Generally, from the above experiment for the two classes, classifying clients as HIV positive outweigh than HIV negative clients. Even though, this model performs better than ever developed so far PART algorithms, it is performing smaller than the one performed by decision tree J48 algorithms in case of method one.

The second method employed in this research to experiments with 10 fold cross validation test options. The result of the four scenario experiment was displayed in Table 5.7.

In the same way, the experiment was made with different scenario for same test option and registered better performance in case of scenario #1 which accounts 96.78% of accuracy than other model. In this method of experiment, the PART algorithm brings better performance in case of the size of tree or rules than in unpruned ones. Because, it reduce the complexity and increase easily understanding of the rules. Additionally, even better performance has been registered than J48 decision tree.

As aforementioned above in J48 decision tree model building, reducing attribute with different test option doesn't bring significance change in the performance of the model rather remains same or very little difference that of all attribute scenarios. This indicates that the reduced attribute (such as SessType, location, causal and sex) has very little or no role in determining HIV infection. The same is true in rule induction learner except minor change in case of scenario with unpruned parameter irrespective of number of attribute. In other hand , those attribute that are selected as the best model has been considered as the major determinate of the HIN infection. From Table 5.6 of method two scenario # 1(PART rule learner with 10 fold cross validation with all attribute) register an accuracy of 96.7% and depicted the predicted result in table 5.8 and see the details confusion matrix in Appendix 3.4.

As shown in table 5.9 below the test result of the confusion matrix the model developed by the PART Algorithm with the 10- fold cross validation, the model scored an accuracy of **96.7**%. This shows that from the total 15396 test data, 14866 (**96.78**%. of the records are correctly classified,

while 530 (3.3%) of them are misclassified. In addition the above table also shows that from the total 7676 HIV negatives clients records 7182 (93.5%) of them are correctly classified as HIV negatives, while 494 (6.5%) of the records are misclassified as HIV positives.

Table 5. 9 Final selected best PART algorithms using 10- fold cross test option

		Predicted HIV infection test result		
		Negative	Positive	Total
HIV infection status (Actual)	Negative	7182	494	7676
	Positive	16	7704	7726
	Total	7198	8198	15396
	10 fold cross validation test option: PART			

Additionally, you can also observe that out of the total 7726 actual HIV positives clients records 7704 (99.79%) of the records are correctly classified as HIV positives clients, while 16(1.21%) of them are incorrectly classified in the HIV negative class. when we compare the HIV positive and negatives class, classifying clients as HIV positive correctly perform better than HIV negative.

To sum up, in the above experiment similarly as of J48 decision tree classifier (method II or 10-fold cross validation) that is conducted using the PART algorithms with all attribute and default parameters values generates a better classification model with a better classification accuracy than 70-30 percentage split test options. So again we selected scenario #1 of 10.fold cross validation model as the best model of PART for classifying clients records as HIV negative and positive and in the mean time we can determine variable that could be risk factor of HIV infection see the snap shot of confusion matrix in appendix 3.4

5.3.1. Feature selection for PART algorithms

One of methods to improve the performance of classification accuracy is implementing feature or attribute selection. In line with J48 algorithms by default has the system automatically select best attribute and generate decision tree only based on best selected feature. For this research pupose the algorithms determines the listed variable mentioned above as the determinate attribute of HIV infection. But PART has no such facility to select the best attribute automatically instead the researcher using default parameter, tried to select feature. The result of this feature includes Residance, Year, Age, MarStatus, EduStatus, ReaHere, PrevTest, EverHSex and instance class are selected out of 20 attribute.

When we compared the performance of the PART algorithms selected as best model running without feature selection and with featured selected running model the former registered 96.7% and the latter 87.2% of accuracy respectively which is a big change or reduction in performance (9.5%). The latter model confusion matrix (**see appendix 3.5**) developed by the PART Algorithm with the 10- fold cross validation, the model scored an accuracy of 87.2%. This shows that from the total 15396 test data, 13426 (87.2.%). of the records are correctly classified, while 1970 (12.8%) of them are misclassified. Additionally, from the total 7676 HIV negatives clients records 6501 (84..7%) of them are correctly classified as HIV negatives, while 1175 (15.3%) of the records are misclassified as HIV positives. And out of total 7720 HIV positive clients records 6925 (89..7%) of them are correctly classified as HIV positive, while 795 (11.3%) of the records are misclassified as HIV negatives..

According to the domain expert evaluation of the selected features are not the only but there is also other determinant variable such as usecodom,employed, steady and HISTI that could be risk factor of HIV infection. As a result the researcher and the domain expert selected the first model (PART algorithms with 10 fold cross validation and all attribute) as the best model that used for predicting HIV infection.

5.3.1. Analyzing Interesting Rules from PART algorithms

It is shown that the best performing model among all models developed using PART was scenario # 1 in Method II illustrated in Table 5.6.The experiment conducted with 20 attribute including instance class, default parameter value (confidence value=0.25,MinNumObj=2 unpruned= false).Running the PART rule induction algorithms on supplied dataset generates rules in plain text. The number of rules generates are 374 and registered accuracy of 96.7%.

To reduce the complexity of the rule and easy understandability the decision rule generated by the above model, the researcher tried by varing the minNumobj(values of 0.2 ,0.25,0.30, 0.40 ,0.50) and minNumObj (value of 2,5,10) are experimented by combining the paired parameter. As a result PART algorithms with parameter confidence factor 0.5 and minNumObj =2 register (96.8%) accuracy which is better than other the selected best PART model. However, the number of rules generated still lengthy so we rely on the first choice.

The partial PART decision list was shown in appendix 5**.** This model generates 121 rules among this some of the most interesting pattern or rules discovered will be illustrates as follows.

1. IF PrevTest = “no” AND EduStatus =” Illiterate “AND LastSex = “no” AND Employed =” no” AND Age = “20-25” then Then the status of client will be HIV positive . Forty five records of the client correctly satisfy this rule. (45.0)
2. IF PrevTest = “no” AND Residence = “Addis” AND Steady =”One partner” AND Year = “3rd year student” AND Category = “staff parent” AND sex = “Female” Then the status of client will be HIV positive. Twelve records of the client correctly satisfay this rule.
3. IF the clients visited voluntary counselling and testing center because of risky and without having sex and whose gender is female then the probability of the client HIV infection is Positive. These rules are correctly found in 34 of the records.
4. IF PrevTest = “yes+ve” AND Employed =” NO” AND UseCond =” Never” AND Age = “20-25” then the status of HIV infection is Positives. These rules are correctly found in 21 of the records.

5. IF Residance = “addis” AND PrevTest = “ yes-ve “ AND Causal = “one partner” AND MarStatus = ” separated” AND HISTI =” No“ AND ReaHere =” plan” AND Employed = “Yes “ AND Age = “20-25” then the status of Client HIV infection is Postives. This rules is correctely found in 18 records.

6. If Residance = “Addis” AND MarStatus = “not married” AND UseCond = “ never”AND LastSex = “ used condom” AND Sex = ”female” AND Age = “20-23”then the status of clent HIV infection is Positive (21.0).

7. PrevTest = “No “AND ReaHere =” causal” AND Year = “3rd year student” AND UseCond= “ never”AND

8. HISTI =” No “AND Age = “” greater than 45 then the status of the client HIV infection is Positive (111.0/3.0)

9. IF Residance = “Addis” AND Loctation = “sadist killo” AND MarStatus = “Married” AND EverHSex = “Yes “AND UseCond = “Never” AND LastSex = “No”AND EduStatus = “other “ ANDReaHere = “risk “AND Age = “>= 46” then Pos (33.0/1.0)

10. If Residance = “addis” AND LastSex =” No” AND EduStatus = ”undergraduate” AND Causal = “one partner” AND Category = “summer student” then client becomes HIV positive (19.0)

11. PrevTest = “yes-ve “ AND Residence = “addis” AND UseCond = “never” AND Causal = “one partner” AND ReaHere =” risk “AND Loctation = “sidist kilo” AND ExpoTime = ” 4to6month” then client most likely HIV positive (14.0)

12. PrevTest = “yes-ve “ AND Residence = “addis” AND UseCond = “never” AND Causal = “one partner” AND ReaHere =” risk “AND Loctation = “sidist kilo” AND ExpoTime = ” 4to6month” then client most likely HIV positive (14.0)

13. IF PrevTest = “No” AND Residance = “Addis” AND Age = “greater than 46 ” AND HISTI = “No “ AND ExpoTime = “less1m “ AND MarStatus =” Marriage” then HIV Positive (22.0/2.0)

14. IF PrevTest = “no” AND ReaHere = “plan” AND EduStatus = “secondary” AND Age = “20.25” Then the status of Clients HIV infection will be negatives. (712.0/4.0)

15. IF Residence = “Addis” AND Steady =”One partner” AND Year = “3rd year student” AND Category = “others” AND Causal =”One partner” AND Age = “greater than or equals to 46” Then the status of client will be HIV positive.(13/5)

16. IF Residence = “Addis” AND Steady = “One partner” AND Year = “3rd year student AND Category = “others” AND MarStatus = “never married “AND Sex = “female Then the probability of Clients HIV Infection status is Positive.(8/1)

17. IF Residence = “Addis” AND Category = “Regular student” AND PrevTest = “No” AND MarStatus = “separated “AND Employed = “yes” Then the probability of the Clients HIV Infection status is Positive.(6/2)

18. IF Residence = “Addis” AND Category = “Regular student” AND MarStatus = “divorced “AND ExpoTime = “no experience” Then the probability of the Clients HIV Infection status is Positive.(11/3)

19. IF Residence = “Addis” AND Category = “Regular student “AND PrevTest = “yes “HIV +ve” AND Sex = “Female” Then the probability of the Clients HIV Infection status is Positive.(10/1)

20. IF Residence = “Addis” AND Category = “Regular student “AND PrevTest = “No” AND EduStatus = “secondary” AND UseCond = “never” Then the probability of the Clients HIV Infection status is Positive.(3/1)

As observed from the above partial list of rules, the classifier has used all attribute to construct rules and provided the class predicted by the model unlike J48 algorithms generated rules. The numeric values which appeared in the bracket next to the class label indicates the number of correctly and incorrectly classified records respectively.

For instance Rule 18 interpreted as clients who came from or live in Addis and registered in regular program and his marital status is divorced and she/he has no experience to exposed to HIV negative then the classifier grouped the behavior under HIV infection status of postive(11/13.0). Meaning in the dataset 1470 records that are exactly satisfied this rule and 17 records indicate misclassified to these rules. Similarly, the remaining rules can be interpreted in the above way.

5.4. Model Building using SMO Algorithms

The third classification task used in this research was Support vector machine. As mentioned in chapter two, the best SVM algorithms that available in WEKA software is sequential minimum optimization algorithms. It works based on risk minimization principles.

To build the model ,even though SMO handles missing value and other noise data automatically the researcher initially clean and handle such issue in chapter four of section data preparation and data cleaning. The same dataset (15396) are used as input for SMO.

The same methods were employed for experimentation as used in the above two models. These are the 10-fold cross validation and the percentage split with 70-30 test option with reduced and all attribute and default values for training and testing the model were used.

Method I: The first experiment of the SVM model building is performed using the SMO algorithm with the default value of percentage split criteria of 70-30 test option. Table 5.10 shows the resulting of two methods and table 5.11 indicate the confusion matrix of the model developed using this algorithm for method one.

Table 5. 10 experimental result of SMO Algorithms with both methods

Experiment	**Scheme**	**# of attribute**	**Accuracy (%)**	**Time (Sec.)**	**CCI**	**ICI**	AVTPR (%	AVFPR (%)	AVROC(%)
Method I: Experiment result with 70:30 percentage Split option									
Scenario # 1	SMO-C1.0-L..	**20**	80.7	2689	3730	889	81	19	81
Scenario #2	SMO -C 1.0…	**16**	80.6	2667	3724	895	81	19	81
Method II: Experiment varying parameter with 10- fold test option									
Scenario #1	SMO -C 1.0….	20	**81**	**2638**	12475	**2921**	**81**	19	81
Scenario # 2	SMO -C 1.0…..	16	80.6	**5463**	12410	2986	97	19	81

As shown in the table 5.10 in method one, we can only observe minor difference in time taken to build a model and remaining experimental result almost similar irrespective of the number of attributes used for each scenario. In this regards, reducing attribute shows only reduce model building time without changing anything else.

So we can generalize that the reduced attribute as mentioned in previous algorithms has no significance for being the risk factor of the HIV infection because as indicate in both method no change at all was made for each model performance. Let me see the confusion matrix of method one scenario # 1.

Table 5. 11 Confusion matrix of SMO with 70-30 percentage split option

		Predicted HIV infection test result		
		Negative	Positive	Total
HIV infection status (Actual)	Negative	1914	426	2340
	Positive	463	1816	2279
	Total	2377	2242	4619
	70-30 percentage split test option: SMO			

As can be seen from the confusion matrix that resulted from the model developed by the SMO algorithms with the 70-30 percentage split, the model scored an accuracy of 80.7%. This shows that from the total of 2340 HIV negatives 1914 (81.7%) of the records are correctly classified as HIV negatives, while 426 (18.3%) of them are misclassified as HIV positives.

Furthermore when we see (appendix 3. 6a) the confusion matrix of the remaining scenario in method two, similarly scenario # 1(Running SMO algorithms with 10 fold cross validation) scored an accuracy of 81 %. This shows that from the total of 15,396 dataset 12475 records are correctly classified as per records of the original dataset and the remaining are misclassified.

The confusion matrix showed in appendix 3.6b out of 7676 total HIV negatives 6318(82,3%) of the records are correctly classified as HIV negatives while 1358 (18.3%) of them are misclassified as HIV positives. In other hand, out of the total 7720 records of HIV positive, 6151(79.7%) are correctly classified as HIV positive and remaining are misclassified as negative.

As one observe from the above Table 5.10 methods II of scenario # 1 is selected the best smo. Meaning the experiment that is conducted with 10-fold cross validation, all attribute and default parameter register best accuracy of 81%.

5.5. Comparison of J48, PART and SMO models

Selecting a better classification technique for building a model, which performs best in handling the prediction and identifying risk factor of HIV infection is one of the aims of this study. For that reason, the three selected classification model with respective best performance accuracy are listed in table 5.12

Table 5. 12 comparison of the result of the selected three models

Types of algorithms	Performance registered (%)	Time taken(sec.)	Correctly classified	Misclassified
J48	**95.5**	1.5	14706	790
PART	96.6	3.37	14866	530
SMO	81	2637	12475	2921
10 fold cross validation test option				

The result showed in table 5.12 that PART rule induction outperforms J48 decision tree and SMO by 1.2% and 15.7% in identifying determinant risk factor of HIV infection. The reason for the PART rule induction to perform better than other is because of linearity of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular HIV infection.

Moreover, in terms of ease and simplicity to the user the PART rule induction is more self-explanatory because it produces in forms of "If then". It generates rules that can be presented in simple human language.

Therefore, it is reasonable to conclude that the PART is more appropriate to this particular case than the J48 decision tree and SMO. So, the model that is developed with the PART rule induction classification technique is taken as the final working classification model to determine or predict HIV infection risk factor to assist both the organization and the university management in case of decision making in matter of HIV related issue.

5.6. Evaluation of Discovered Knowledge

Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006).Besides, data preparation task is a crucial so that data mining tools can understand the data.Because without doing so the accuracy of the discovered knowledge will be poor.

Thus, data cleaning and analysis methods and outlier mining methods becomes a must to tackle such gab before discovered knowledge so as to enhance the efficiency and effectiveness of the data mining technique. In this effort, the researcher tried to assess such problems which require due emphasis while making the data ready for applying data integration and other different data reduction techniques. See the details effort that has been done in this regards in chapter four.

As discussed in chapter two and five, in selection of model building, three classification models was selected and developed for reason explained below. Firstly, the J48 decision tree algorithm

is chosen for is simplicity easy interpretability of the result generated by it. Secondly PART is selected similarly as of J48 and simplicity and generated list is easy to interpret and associate the problem domain. Finally, because of the recommendation of previous research studied in area of HIV/AIDS the researcher selected SMO algorithms. Above all, three of the model was experimented, evaluated and compared as a result with 10 fold cross validation test option , PART rule induction learner was used as a final model for study.

From 374 all possible list of decision rules generated only few of the rules are discussed. These rules are unambiguous, easy to interpret and understand. see more about PART decision rule set on Appendix 5.

As illustrated in appendix 5, all attributes are used for construction of decision rule set. As indicated in section 5.3.2 experiment has been implemented by applying WEKA feature selection technique. However the performance is less than those implemented with all attribute .As a result of this research output is evaluated based on all attribute generated decision rule set. Hence, the most interesting rules or discovered knowledge were:

Rule 1: IF Residence = “Addis”AND Category = “Regular student” AND MarStatus = “divorced “AND ExpoTime = “no experience” then HIV Positive.(11/3)

Rule 2: IF ReaHere=”risk” AND ExpoTime =”no experience” AND Sex=”Female”: Postives(34.0)

Rule 3: IF PrevTest = “yes-ve “ AND Residence = “addis” AND UseCond = “never” AND Causal = “one partner” AND ReaHere =” risk “AND Loctation = “sidist kilo” AND ExpoTime = ” 4to6month” then client most likely HIV positive (14.0)

Rule 4: IF PrevTest = “no” AND EduStatus =” Illiterate “AND LastSex = “no” AND Employed =” no” AND Age = “20-25”: postives (45.0).

Rule 5: IF PrevTest = “yes+ve” AND Employed =” NO” AND UseCond =” Never” AND Age = “20-25”: Postive(21.0)

Rule 6: IF Residance = “addis” AND PrevTest = “ yes-ve “ AND Causal = “one partner” AND MarStatus = ” separated” AND HISTI =” No“ AND ReaHere =” plan” AND Employed = “Yes “ AND Age = “20-25” : Positive(18)

Rule 7: PrevTest = “No “AND ReaHere =” causal” AND Year = “3rd year student” AND UseCond= “ never”AND HISTI =” No “AND Age = “greater than 46” : positive(111.0/3.0)

Rule 8: PrevTest = “No “AND ReaHere =” causal” AND Year = “3rd year student” AND UseCond= “ never”AND HISTI =” No “AND Age = “” greater than 45 then the status of the client HIV infection is Positive (111.0/3.0)

Rule 9: IF Residance = “Addis” AND Loctation = “sidist kilo” AND MarStatus = “Married” AND EverHSex = “Yes “AND UseCond = “Never” AND LastSex = “No”AND EduStatus = “other “ AND ReaHere = “risk “AND Age = “>= 46” then Pos (33.0/1.0):

Rule 10: IF PrevTest = “no” AND ReaHere = “plan” AND EduStatus = “secondary” AND Age = “20.25” Then the status of Clients HIV infection will be negatives. (712.0/4.0)

Rule11.If PrevTest=”yes HIV +ve” AND ReaHear=”other”AND ExpoTime=“donot known” AND MarStat= “never married” LastSex = “yes”Then HIV negative.

Rule 12 IF Residence = “Addis” AND Category = “Regular student “AND PrevTest = “yes “HIV +ve” AND Sex = “Female”: Positive(10/1).

The rules that are accessible above indicate the possible conditions in which VCT service clients record could be classified in each of HIV Negative and positives classes. All attributes are participated in construction of decision rule set but as observed above the contribution in each of the experiment and domain expert validation four of the attribute (Sesstype, sex, location and steady) are less determinate of HIV infection.

From the above list of partial PART rule induction generated decision rule set or discovered rules/ knowledge, we can observe how the relation of attribute contribute towards HIV infection.

A client more likely to be HIV negative even if previous test is HIV positive. Rule 11 is a good indicator of this fact. The domain expertise will also agree with this finding because it happens because of malfunction of the laboratory or some man made problems. Other hand the laboratory result of the client may be HIV positives even if previous laboratory test of a client said HIV negative. This because the virus is not visible within window period (meaning a client can confirm his test result after three month). Rule 4, 7 and 8 shows this relationship. The domain expert also supported this idea by saying clinically approved situation.

Moreover clients whose previous test was HIV positive (Rule 5) might not be change for second time laboratory result. This is for the reason that clients have been infected and cannot recover from diseases rather prolong their life span by taking “tablet” for limited year. Furthermore, Clients whose previous HIV status is negative have also been shown as negative during repeat testing (Rule 10). This may reflect that there is some behavioral change after getting VCT service. UNFPA (2002) reported that understanding status of HIV negative can serve as a strong motivating factor to remain negative and is considered as evidence from the existing knowledge in the domain.

According to Rule 7 and 9 age group above 45 years old are more exposed to HIV infection. For instance, Rule 9 that client who comes from Addis and currently found in sadist kilo campus whose age was greater than 45. Additionally who got married and never use condom and their educational status is unmentioned; they came in VCT center because of something that they consider risk for HIV infection then the client is suspected to HIV infection. Based the classification accuracy of the PART algorithms 33 of the record dataset is implied this. In this regard the domain expert reported that this situation is true sometimes it occurred because, more attention is given through different media on productive age (age from 20-30). Health professional do not always test older people for HIV/AIDS and so may miss some cases during routine check-ups. Additionally, sometimes people above 50 often mistake signs of HIV/AIDS for the aches and pains of normal aging, so they are less likely than younger people to get tested for the disease.

Furthermore, this research finding indicates that those people whose ages between 20-25 are more exposed to the HIV infection. For instance, Rule 4, 5 and 6 reported the effect of this rule. To elaborate clients who never used condom and their age is among 20 and 25 irrespective of employment status may be more susceptible to HIV infection. The USAID (2006) report has also stated that since AIDS deaths are concentrated in the 25 to 45 age group, communities with high

rates of HIV infections lose disproportionate numbers of parents and experienced workers and create gaps that are difficult for society to fill.

In real environment and many research suggest, females are more susceptibility to HIV Infection than male (UNAIDS, 2000). This is indicated by Rule 2 and 12. Rule 12 report that those clients whose residence was Addis and registered for regular program having job and knows her status as HIV positive earlier would be classified as HIV positive. Moreover, that female client that has no experience to susceptible to to HIV infection and consult VCT center for reason of risky condition is vulnerable to HIV infection.

Lastly, this research finding showed that clients whose level of education or year in which they achieved as 3rd year university student are exposed to HIV infection. Rule 7 and 8 reported this situation. The domain expert explained this condition is true. Additionally, in real environment of university freshman students are more exposable than senior student. This is because most of these students are new to university and exposed to various thing such as chewing chat, drinking and smoking cigarette. This is the major factor of HIV infection in university (Negatu and Seman, 2011). By doing so, they can be easily exposed to the diseases.

In general, the domain expertise also agree with the general idea that the model produced and adds some fact that some clients also vulnerable as a result of addiction which has not included as a behavior of clients in dataset. So having knowledge of their addiction of the client is very important to determine and predict their status.

The finding interpreted above age, previous test, cause, exposure time, marital status, last time use of condom in sexual intercourse, use of condom, sex, educational status, employment status, residence, category, year and reason for taking VCT services are the major determinant factor of HIV infections of clients.Some of the attributes (Location and sesstype) that the algorithms assumed less important also less important for domain expert in determining HIV infection status of this study.

The classification model used in this research to predict as risk factor of HIV infections are pretty good .They come up with reasonable accuracy performance. Among this model PART rule learner register better than other and an accuracy of 96.7 %.

The overall experimentation process of experimentation has been performed following the diagram.

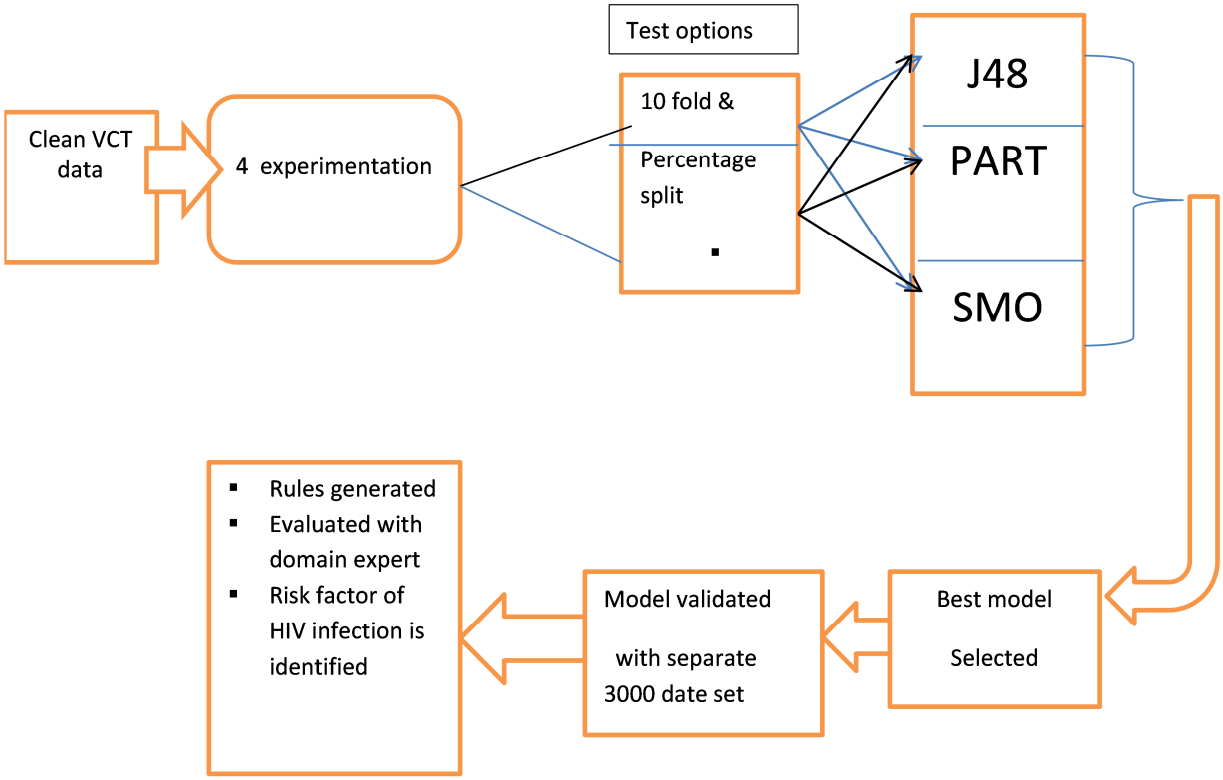


Figure 5.4 Graphical representation of Experimentation performed in this thesis

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

A major challenge facing healthcare organizations (VCT Service Provider, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable (Srinivas et al., 2010).Besides, health care centers including VCT service provider must also minimize the cost of clinical tests.

To this end, they can achieve these results by employing appropriate computer-based information and/or decision support systems particularly application of data mining technology. Health care related data including VCT data is massive. Hence, such data must be analyzed by organization using data mining technology. Hence application of data mining techniques may help in answering several important and critical questions by extraction of useful knowledge that enables support for cost-savings and decision making.

In this research, an attempt has been made to apply the data mining technology in support of identifying and predicting HIV infection risk factor using VCT data. Data mining technology basically follows an iterative process consists of: Data collection, Data preparation and understanding, model building and evaluation. As discussed different data mining methodology in chapter two , there is a possibility of additional steps to above list for instance KDD includes nine steps. The iterative nature of the process assist the data miner to back and forth at different and fix it where problem arise.

Accordingly, this research used and followed the six steps Cios et al (2007) methodology to investigate the stated problem. The methodology has strictly been followed while undertaking the experimentation. The data used in this research has been gathered from African AIDS

Initiatives International VCT center. Once the data has been collected (15396 dataset), it has been preprocessed and prepared in a format suitable for the DM tasks. Data preparation took considerable time of the study.

The study was conducted using three classification techniques namely decision tree, rule induction and support vector machine. For model building and experimentation J48, PART and SMO algorithms are used. Experimentation was conducted using four scenarios in two methods for each algorithm except for SMO. For SMO only two scenarios were used for the two methods with default value and the output of this is used as a comparison for former two algorithms.

By changing the training test options and the default parameter values of the algorithm, these models are tested and evaluated. As a result PART rule induction algorithms registered batter performance with 96.7 % accuracy running with 10-fold cross validation with default parameter using 20 attribute than any experimentation done for this research purpose. The performance of the model is validated using randomly selected 3000 datasets and registered performance accuracy of 95.8% which is pretty good model.

According to the result of PART algorithms, all the 20 attributes are used to construct the decision rule set. Besides, using WEKA feature attribute selection is tried and tested the performance is 77 %. This implies that decrease in performance resulted because of missing of relevant attribute. To this end, with extensive discussion of domain expert and “educated guesses” out of the 20 attribute, only the following attributes are selected as the determinate variables of HIV infections has been identified.

These are previous test of HIV, use of condom, Reason hear in VCT center, History of sexual Intercourse, employment status, HIV infection causal situation, Last sex with condom, exposure time for risk situation (number of partner the client faced), marital status, educational status, year, age .

Besides that studying the behavior of the university community related to addiction such as chewing chat, smoking cigarette and drinking are important to identify the risk factor of HIV infection.

Moreover, the finding of this research indicates that females are more vulnerable than men to HIV/AIDS. Third year university students are more susceptible to HIV infection. Previous test of client who have HIV negative and for second time test is negative. The laboratory result may be may be positive because the virus may not be visible until end of three month. A client who has got HIV positive may be result in second test "negative "which is because of health professional or malfunctioning of laboratory test or equipment.

Analogously, J48 and SMO algorithms were also register an encouraging performance of classification accuracy. Both model correctly classify new instance of VCT clients as 95.5 % and 81 % with 10 fold cross validation with all attribute and default parameter. When we compare the result of three model developed based on accuracy registered to correctly classify new instance of VCT data, PART perform better.

In general, in this study only few experiments are carried out using J48, PART and SMO with few parameter setting. Missing values are simply replaced by most frequent values using WEKA 3.7.5 and manually for few attribute in consultation of domain expert which may sometimes create bias so that if time allows it is better to use other mechanism like filling extensively by consultation of domain expert. Furthermore, the researcher faced challenging in experimentation of support vector machine because of time consuming, balancing of the unbalanced dataset and getting of VCT dataset.

In spite of the challenges and weaknesses in data pre-processing and limited application of algorithms the results are encouraging. The results obtained from this research indicate that data mining is useful in bringing relevant information to the service providers as well as decision makers.

6.2. Recommendation

In this research work an attempt has been made to find out the potential applicability of data mining technology to support the prevalence of HIV activity at Addis Ababa University community and surrounding based on the data accumulated on VCT service by AAII. Even though, this research has been done for the academic exercise its results are found promising to be applied in addressing practical problems in prevention and control of HIV /AIDS.

Hence, based on the findings of this research, the following recommendations are forwarded:

- This research has proven the applicability of PART algorithms which automatically discover hidden knowledge that are interesting and accepted by domain expert. However, this research output doesn't integrate with knowledge based system. To do so, the researcher recommends implementing the discovered classification rules with domain knowledge as a knowledge based system that could be helpful for health professional, epidemiologists and other researchers who are conducting HIV/AIDS care, control and prevention.

- Attribute such as session type and campus were not good discriminate variables for predicting HIV infection, hence to investigate further research in case of prevalence of HIV in higher education institute, these attributes will not be recommended

- This research has attempted to determine HIV infection risk factor using classification data mining task only. The association that exists between the selected attributes related to HIV infection was not considered. Therefore, research can be conducted to see the degree of association that coexists between those attributes and HIV infection is forwarded.

- For this work, balanced dataset is used to determine the risk factor of HIV infection. However, researches can also be conducted using actual imbalanced data and other data mining technique such as Class Confidence Proportion Decision Tree (CCPDT). This is a robust decision tree algorithm for imbalanced datasets.
- This research has been attempted to determine the risk factor of HIV infection with limited data set and 20 attributes collected from Addis Ababa university community and its surroundings. Further researches can also be conducted by increasing the number of attributes and datasets from all higher education institutions in Ethiopia.
- The research also attempted to test the applicability of data mining in identifying HIV infection risk factor. In doing so, implementing using more data mining techniques such as Naïve Bayes and Multilayer perception may boost the performance of the model and helps to compare the result with this research finding.

REFERENCESE

AAU (2006).HIV/AIDS Policy for Addis Ababa University (AAU).Prepared under the joint Auspices of the Addis Ababa University and African AIDS Initiatives International Inc. Addis Ababa,Ethiopia:AAU press

AAU to Host 60th Graduation Day (AAUa). Press release July 2011 .Rretrieved on March 2012 from www.aau.edu.et

Abdinasir A, Gurja B, Sentayehu T (2002). Knowledge, Attitude, Practice and Behavior (KAPB): Survey on HIV/AIDS among AAU Students, AAU, Ethiopia: UNFPA, *ISAPSO*.

Abdul-Kareem,S., Baba.S. , Zubairi, Y. Z., Wahid, M. I. A. (2000). Artificial Neural network (ANN) as a tool for medical prognosis. Sage publication

Abraham Tesso (2005) Application of Data mining Technology To identify the Determinate risk factors of HIV infection and To find Their association Rules: A Case of Center for Disease control and Prevention. Master Thesis. Addis Ababa University: School of Information Science.

Addis Ababa Highlights (AAH) (2006) magazine issue vol.3.(2)Retrieved March 2012 from www.unep.org/roa/Addis_Ababa_Site/.../February2006.pdf

Africa AIDS Initiative International (AAII) (2006). HIV/AIDS, STDS and Reproductive Health: Knowledge, Attitude, Practice and Behavior (KAPB) survey on AAU Students, Final Survey Report Addis Ababa, unpublished.

Agrawal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rule in large databases', San Jose, California: IBM Almaden Research Center

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996) Fast discovery of association rules: In Advances in Knowledge Discovery and Data Mining, pages 307–328. AAAI Press

AIDS (n.d.) All about AIDS Discovery. Retrieved On march 2012 URL: http://library.thinkquest.org/03oct/01335/en/txt/discovery.html

Behavioral Surveillance Survey (BSS) Ethiopia (2005). Round Two. AddisAbaba, Ethiopia : MOH/HAPCO, AAU, CSA, EPHA

Berry, M. & Linoff , G. (2004). Data mining techniques for marketing, sales, and customer relationship management. (2nd ed.).Indiana: Wiley publishing.

……. (2000) Data Mining Techniques for Marketing Sales and Customer Support. New York: John Wiley & Sons.

Berry, Michael J.A. and Gordon Linoff(1997). Data Mining Techniques For Marketing Sales and Customer Support. New York: John Wiley & Sons.

Berson ,Alex, Smith ,Stephen, and Thearling ,Kurt(n.d.) An Overview of Data Mining Techniques Retrive on Feb, 2012.From http://www.thearling.com/text/ dmtechniques/dmtechniques.htm

Beyene P, Solomon B, Yared M (1997). AIDS and college Students in Addis Ababa: A Study of Knowledge, Attitude and Behavior. Addis Ababa ,Ethiopia. : J. Health Dev. 11(2):20-30.

Bharti, K Jain, S and Shukla, S (2010) Fuzzy K-mean Clustering Via J48 for Intrusiion Detection System.International Journal of Computer Science and Information,1(4).

Birru Asmare(2009). Application of Data mining Technology to support VCT for HIV: A Case of Center for Disease control and Prevention. Master Thesis .Addis Ababa University. School of Information Science.

Cabena, P., Stadler,R., Hadjinian ,p. , Zanasi, A., and Verhees,J.(1998). Discovering data mining from concept to implementation. New Jersey: Prentice Hall

Chakrabarti ,s, Cox,E, Frank ,E., Hartmut ,G.R., Han, J. Jiang,X. Kamber,M and, Witten, Ian H.(2009)Data mining: know it all. Burlington. San Francisco, CA, USA: Morga Kaufmann Publishers

Chapman ,P., Clinton, J., Kerber ,R., Khabaza, T., Daimlerchrysler ,T. R., and Shearer ,C(.2000) "CRISP-DM 1.0: Step-by-step data mining guide." CRISP-DM consortium. SPSS.USA

Chawla, Nitesh V.(2003)." C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure". Workshop paper on Learning from Imbalanced Datasets II, Washington DC: ICML,

Cios, K ,Witold, P, Roman, S and Kurgan ,A. (2007). Data Mining: A Knowledge Discovery Approach, New York, USA: Springer,

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. Machine Learning,.3(4):261-283.Avilable at www.citeulike.org/user/janch/article/9304405.

Daniele M. and Satheesh R(n.d) . Improving tax administration with data mining, Elite Analytics, LLC. Retrived march2012 from http://www.spss.ch/file.php?file=/upload/1123070263_Improving%20tax%20administration

Deshpande, S P, and V M Thakare.(2010). Data mining system and applications : a review. International Journal 1 (1): 32-44.

DMSHO (2011).Data mining for Successful Healthcare Organizations (DMSHO)Retrived March 2012 from http://www.shamsgroup.com/Portals/0/Articles_in_the_media/DataMining.

Dunham, M. H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

Eapen,A.G.(2004) Application of Data mining in Medical Applications.Master Thesis . University of Waterloo. Waterloo, Ontario, Canada.Retrived on April 2012from

Elias Lemuye(2011).HIV Status Predictive Modeling Using Data Mining Technology. Master Thesis .Addis Ababa University.School of Information Science and Public Health. Unpublished

Elias Worku (2009). Internet –Based Technology as source of sexual and HIV/AIDS Related Health Information among AAU students. Unpublished Thesis

Famili, A and Turney, P.(1997). Data preprocessing and Intelligent Data analysis. Institute of Information Technology, National research council, Canada

Family Health International (FHI) (2004).HIV Voluntary Counseling and Testing: A Reference Guide for Counselors and Trainers. Retrived on March 2012 from URL: www.fhi360.org/en/HIVAIDS/pub/guide/vcttoolkitref.htm

Fayyad, Usma, Piatetsky-shpiro, G. and smyth, padharic.(1996). From Data Mining to knowledge Discovery in Databases. Available URL: http://citeseer.nj.nec.com/fayyad96from.html

FDRE HIV/AIDS PC (2010). FDRE HIV/AIDS Prevention and Control Office. Report on progress towards implementation of the UN Declaration of Commitment on HIV/AIDS Retrived On March 2010. From: http://www.unaids.org/en/dataanalysis/monitoring countryprogress/2010progressreportssubmittedbycountries/ethiopia_2010_country_progress_report_en.pdf , Accessed on December,2011

Frank,E, and Witten ,Ian .(1998).Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning,144-151,

Garbus L (2002) HIV/AIDS in Ethiopia, Country AIDS Policy Analysis Project, San Francisco: AIDS Policy Research Center, University of California

Getinet T. (2009). Self-reported sexual experiences, sexual conduct and safer sex practices of AAU undergraduate male and female student's on the context of HIV/AIDS pandemics, A PhD dissertation Buffalo, USA: the State University of New York

GHIVR (2011) Progress report 2011: Global HIV/AIDS Response(GHIVR). World Health Organization HIV/AIDS Department, Switzerland, Geneva.

GHSSH (2011).Global Health Sector Strategy on HIV/AIDS(GHSSH)2011–2015. Geneva,World HealthOrganization, 2011 Retrived on March 2012 from (http://whqlibdoc.who.int/publications/2011/9789241501651_eng.pdf, accessed March 2012).

Government of Ethiopia GoE (2004) A Comprehensive Strategic Plan to Combat HIV/AIDS Epidemic in Ethiopia (2004 – 2007), Final Report, Addis Ababa

Grzymala-busse, jerzy w.(n.d.) Rule Induction. University of Kansas Retrieved on February,2012.From people.eecs.ku.edu/~jerzy/c95-perugia.pdf

Gupta, s., Kumar, d. and Sharma, A.(2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis . Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 (2).

Hailom, B., Aklilu, K. , Sara, B. and ,Kate, S.(2005).The System-Wide Effects of the Global Fund in Ethiopia Baseline Study Report. Maryland,Abt Associates Inc. Retrieved from URL: www.theglobalfund.org/.../library/Library_IEPNADF194_Report_en.March 2012

Han, J & Kamber, M. (2001).Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

---------- (2006).Data mining: concepts and techniques. (2nd ed.). San Francisco: Morgan Kaufmann Pubisher

Hipp, J. and Ulrich, H. (2000) 'Algorithms for Association Rule Mining- A General Survey and Comparison', Tubingen: University of Tubingen

UNAIDSHIV) (2000).HIV testing methods (UNAIDSHIV): UNAIDS Technical Update. Geneva, UNAIDS. Retrieved March 2012 From URL: http://data.unaids.org/Publications/irc-pub01/jc379-vct_en.pdf

Hogg, D. W., Jackson, C., ˙Zytkow, A. N., Irwin, M., Webster, R., and Tremaine,H(2006). Rates of disease progression according to initial HAART regimen: A collaborative analysis of 12 prospective cohort studies. Journal of Infectious Diseases 194: 612-22

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning. 11:63-91.

Jinhong, L Bingru, Y and Wei, S (2009). A New Data Mining Process Model for Aluminum Electrolysis. Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA'09) Qingdao, P. R. China : Academy Publisher. Pp.193-195.

Kantardzic ,Mehmed(2003). Data mining: concepts, models, methods, and algorithms. The University of California; Wiley-Inter science

Kathy, L. and Stacey.(2009). From the Ground Up: User Involvement in the Development of an Information System for use in HIV Voluntary Counselling and Testing. Retrieved on March 2012fromURL:http://cit.mak.ac.ug/iccir/downloads/ICCIR_09/Kathy%20Lynch%20and%20Stacey%20Lynch_09.pdf

Koh, Hian C. and Tan, Gerald (2005). Data mining applications in Healthcare. Journal of Health Information Management,19(2) Available at: http://www.himss.org/content/files/Code%20109_Data%20mining%20in%20healthcare_JHIM_.pdf accessed on April 2012

Kohavi ,Ron(1995). The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189,

Kurgan and Musilek (2006) A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, Vol. 21:1, 1–24. 2006, Cambridge University Press

Larvac, Nada.(1998) Data Mining in Medicine : Selected Techniques and Applications. Retrived On April 2012 from URL.: http://citeseer.nj.nec.com/lavrac98data.html

Lee,W., Stolfo, S.J., and Mok, K.W. (1999). A Data Mining Framework for Building Intrusion Detection Models. In IEEE Symposium on Security and Privacy (pp. 120–132).

LLGP (2007). Lessons Learned and Good Practices on SRH and HIV/AIDS Prevention: Summary of Project Performances of 12 Local NGOs Supported by UNFPA/NORAD. Unpublished.

Lloyd -Williams, Michael. (1997).Discovering the Hidden secrets in your Data the data Mining approach to Information. Retrieved March 2012 from URL:http://informationr.net/ir/3-2/paper36.html

Madigan ,Elizabeth, Curet, O. and Zrinyi, Miklos (2008). Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns.Human Resource for Health licensee BioMed Central Ltd.,6(2).1-6

Mamcenko, Jelena and Beleviciute. Inga (2007). Data Mining for Knowledge Management in Technology Enhanced Learning. Journal of Knowledge Management: P.115-119.

Matsatsinis ,N. F., Ioannidou, E. ,and Grigoroudis ,E..(n.d.)Customer satisfaction using data mining techniques Chania. . Technical University of Crete.

Ministry of Health (MoH)(2000).Health and Health Related Indicators, Planning and Programming Department, MoH.unpublished

Ministry of Health (MOHHA, 2006) - HAPCO. Accelerated access to HIV/AIDS prevention, care and treatment inEthiopia. Road map 2007-2008/10. Accessed on March 2012 from URL: http://fitun.etharc.org/resources/healthpolicyandmanagementresources/roadmap2007-2010ethiopia.pdf

Ministry of Health(MOH) (2006).AIDS in Ethiopia. Fifth Report. Disease Prevention and Conttol Department, MOH.

Moshkovich,Helen M., Mechitov,Alexander I., and Olson,David L. (2002) .Rule induction in data mining: effect of ordinal scales. Expert systems with applications. Elsevier Science Ltd. . Technical University of Crete. 22 p.303-311,

Mung'ala, L., Taegtmeyer ,M., Kilonzo, N., Morgan, G. and Theobald ,S.(1995).A community-based counselling service as a potential outlet for condom distribution. 9th International Conference of AIDS and STD in Africa. Kampala, Uganda.

My Data Mining Weblog (MDMW).(2008).Rule learner (or Rule Induction) Retrived Febrary 2012 From: http://mydatamine.com/?tag=induction-algorithms

National AIDS Council (NAC)(2001)Strategic Framework for the National Response to HIV/AIDS in Ethiopia. (2001-2005).

Nigatu and Seman (2011).Attitudes and practices on HIV preventions among students of higher education institutions in Ethiopia: The case of Addis Ababa University.Available at http://interesjournals.org/ER/pdf/2011/February/Nigatu%20and%20Seman.pdf accessed October,2011

Obenshain, Mary K.(2004). Application of data mining techniques to healthcare data.: Infection control and hospital epidemiology .The official journal of the Society of Hospital Epidemiologists of America 25 (8): 690-5.

Pal, N.R., Jain, L.C., (2005).Advanced Techniques in Knowledge Discovery and Data Mining, Verlag:Springer

Palaniappan and Awang (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. IJCSNS International Journal of Computer Science and Network Security,8 (8)

Pete, C. (2000).Step by step Data mining guide. CRISP-DM consortium,.6: 60-67

Piatetsky-Shapiro, Gregory.(2000) .Knowledge Discovery in Databases: 10 Years After. SIGKDDExplorations.1Online.Internet. Retrived on march 2012 http://www.kdnuggets.com/gpspubs/sigkddexplorations-kdd-10-years.html

Plate, Bert,J, Grace,J, and Band(1997).A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer. Advances in Neural Information processing :MIT press retrived on April 2012 from http://citeseer.nj.nec.com/plate96comparison.html

Platt, John (1998).Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. MIT Press.

Popa, H., Pop, D., Negru, V. & Zaharie, D. (2007). AgentDiscover: A multi-agent system for knowledge discovery from databases. In Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing.

Prather, M. Dentener, F. , Derwent, R. , Dlugokencky, E. , Holland, E. , Isaksen, I. , Katima, J. ,Kirchhoff, V. and Matson, P.(2001)Medical Data Mining : Knowledge Discovery in a clinical Data Woehouse. Retrived on April 2012 fromURL: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput605/2008/pdf/medical_data_mining_prather.pdf Accessed on October,2011

WKDD (2008) . Advanced Knowledge Discovery and Data Mining. IEEE discovery and Data Mining, Workshop on Knowledge Discovery and Data Mining. University of Adelaide.; IEEE Computer Society.

Quinlane , J.R. (1986). Induction of Decision Trees. Kluwer Academic publisher. Boston (1), 81-106

Raghavan, P. , Manning, C.D. , Schütze, H(2002) .Data Mining Trends and Issues. Retrived On April 2012 from: http://citeseer.nj.nec.com/138316.html

Razak,A. and Bakar,A. (2006). Association Rules Mining in Asthma Patients Profile Dataset. Chiang Mai Journal of Science, 33(1): 23-33 (ISSN: 0125-2526).

Rea, Alan (1995). Data mining an introduction Version 2.0, Queens University Belfast.)].261– 283.

Refaat, Mamdouh (2007).Data Preparation for Data Mining Using SAS. sanfransisco: Mamdouh Refaat publisher

Roiger, Richard. And Geatz, Michael (2003). Data mining: a tutorial-based primer.Addison Wesley,

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C 35 (4): 476–487.

Santos , Manuel Filipe,& Azevedo,Ana (2008).KDD, SEMMA AND CRISP-DM: a parallel overview .. IADIS European Confrence data miing.p.182-187.

Schneider H et al(2010). Urban–rural inequalities in access to ART: results from facility based surveys in South Africa. AIDS 2010 – XVIII International AIDS Conference, Vienna, 18–23 Retrieved March 2012 from URL: http://www.iasociety.org/Default.aspz?pageid=12&abstracted=200738997).

Sheikh, Khalid (2008). Using Decision Trees to predict customer behavior. Indian Expressed group. Retrieved on May 2012 http://expresscomputeronline.com/20040412/technology01.shtml

Shermer, Michael (2002).The Skeptic Encyclopedia of Pseudoscience 2 volume set.ABC-CLIO., California

Silltow, John(2006). Data mining: Tools and Technique. Retrived on March 2012

From at http://www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/

Smolander, K., Tahvanainen, V.P., Lyytinen, K., (1990) How to Combine Tools and Methods in Practise - a Field Study. In: Lecture Notes in Computer Science, Second Nordic Conference CAiSE'90, Stockholm, Sweden. pp. 195-211.

Srinivas, K. et al. (2010) Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSEI) Vol. 02, No. 02, 250-255

Tefera B, Challi J, and Yoseph M (2004). KAP about HIV/AIDS, VCT Among students of Jimma University, Jimma Zone, SW, Ethiopia. Ethiop. J. Health Sci. Vol.14, Specal issue.

Teka T (1993). College Students' Attitude and knowledge of AIDS. Ethiop. Med. J. 314:233-237.

Tekelu(2010). Application of Data mining Technology on Antiretroviral Therapy(ART) data: a case of Adama and Assella Hospital. Master Thesis .Addis Ababa University. School of Information Science

Trybula, Walter J.(1997) Data Mining and Knowledge Discovery. ed. Williams, Martha E. Annual Review Information Science and Technology., Vol. 32, Information Today, Inc. New Jersey, USA: Medford,

Two crows Corporation (1999), Introduction to data mining and knowledge discovery (2nd Ed).by Edelstein, H., A. Potomac, MD: Two Crows Corp.

…… (2005) Introduction to data mining and knowledge discovery (3rd Ed).by Edelstein, H., A. Potomac, MD: Two Crows Corp.

USAID (2006). Bringing information to decision makers for global effectiveness:how HIV and

AIDS affect population. Population Reference Bureau: Washington.

UNAIDS (2010). Global report: UNAIDS report on the global AIDS epidemic. Vol.2007, issue, Geneva.

UNAIDS(2011).World AIDS Day Report of 2011. Geneva, UNAIDS.Retrived From: www.unaids.org/.../unaids/.../unaidspublication/2011/JC2216 Accessed on March 2012.

UNAIDS/WHO (2010). AIDS epidemic update UNAIDS 20 avenues Appia CH-1211 Geneva 27 Switzerland.

UNAIDSa(2011). The Global HIV/AIDS Epidemic: Fact Sheet, 2011. Geneva, UNAIDS. Retrieved from: www.kff.org/hivaids/upload/3030-16.pdf .Accessed March, 2012.

UNAIDSb(2010). Report on the Global AIDS Epidemic; Geneva, UNAIDS.Retrieved On March 2012 URL: www.unaids.org/globalreport/global_report.htm

UNAIDSc(2011). AIDS at 30: Nations at the crossroads; 2011.Retrieved On March 2012 URL: www.unaids.org/unaids_resources/aidsat30/aids-at-30.pdf

USAIDHPI (USAID Health Policy Initiative). Equity: quantify inequalities in access to health services and health status. Washington, DC, Futures Group, Health Policy Initiative, Task Order 1, 2010 (http://www.healthpolicyinitiative.com/Publications/Documents/1274_1_ EQUITY_Quantify_FINAL_Sept_2010_acc.pdf, accessed March 2012).

UNFPA (2002).Voluntary counseling and testing (VCT) for HIV Prevention. HIV Prevention Now Program Brief 5: 1-2.

Vararuk, A., Petrounias , I. , & Kodogiannis, V.(2008).Data mining techniques for HIV/AIDS data management in Thailand. Journal of Enterprise Information Management,21(1):52-70.

Witten, I. H., & Frank, E. ,2000 ,"Data mining concepts" ,New York, Morgan-Kaufmann

Wong, M. L., Leung, K. S.(2000). Data Mining using Grammar Based Genetic Programming and application. Springer

World Health Organization (WHOa).(2010) Towards Universal Access: Scaling up priority HIV/AIDS interventions in the health sector. Progress Report, September 2010. Retrieved March 2012 http://www.who.int/hiv/pub/2010progressreport/summary_en.pdf

World Health Organization (WHOb) (2004). Policy statement on HIV testing. Retrieved on March, 2012 URL:http://www.who.int/hiv/pub/vct/statement/en/index.html

World Health Organization (WHOS) (2005).Scaling-up HIV testing and counselling services: a toolkit for program managers. Geneva, Switzerland

Appendix 1: Standard Format Of AAII VCTUser Registration Form

Appendix 2: Descriptions of Selected Attributes

No.	Attributes	Descriptions	Codes & Corresponding Values
1	Age	Age of the clients	{A1=0-15, A2=15-19, A3=20-25,A4 =26-35, A5=36-45,A6 >= 46}
2	Sex	Sex of the clients	{Male=M, Female=F}
3	SessType	Session type of the clients	{ indv=Individual , Coup=Couple , other=Group}
4	MarStat	Marital status of the clients	{ marr=Married , Nmar=Never married , separ=Separated , divor =Divorced , wido=Widowed , liv2= Living together ,
5	Education	Educational level of the clients	{Illitrate=Illiterate , read=Able to read , prsch= Primary , sesch= Secondary , ugrad=under grad,pgrad=postgraduate
6	Employed	Employment status of the clients	{N=No , Y=Yes }
7	ReaHere	Primary reason the clients visit the center	{Risk , Plan ,status, causal other}
8	PrvTeste	Whether the clients are previously tested	{ No=No ,Yes+ve=Yes HIV positive , “Yes-ve”=Yes HIV Negative , Yes-inc =Yes inconclusive ,NA=Didn’t take
9	EverHSex	Whether the clients have ever had sex	{no=No , yes=Yes}
10	UseCond	Condom usage behavior of clients during sexual intercourse	{Ne=Never , Al=Always, so=Sometimes, Not aplicable=Na }
11	LastSex	whether clients have used condom the last time s/he had sexual intercourse	{N=No , Y=Yes, DN=Doesn’t remember, 99=NA }
12	HistSTI	client’s lifetime history of diagnosed or suspected Sexually Transmitted Infections history	{N=No , yeno=Yes , dnknow=Don’t know , 98=NA}
13	Casual	Number of all those sex partners the client has seen only once or rarely for the past 3 months	{np=No Partner , 1p=One Partner, mp=Multiple Partners}

14	Steady	Number of all those sex partners that the client considers as regular partners for the past 3	{np=No Partner , 1p=One Partner, mp=Multiple Partners}
15	Year	Duration of client in attending lecture successfully in university	{Y1=1_{st} year, Y2= 2_{nd} year ,Y3=3^{rd} year,Y4=4 th year, Y5 fifth year,,Y6= year six , Ny=no year}
16	ExpoTime	Suspected Exposure Time to HIV	{less1m = < 1 month , 1mto3m = 1 to 3 months , 4to6m = 4 to 6 months , greater6m= over 6 months , noexpo =
17	Residence	Clients who came from to the university	Region 1=R1,Region 2=R2…….R14 and No region=other
18	Location	The place where the client is currently living	{Com= commerce, A= not applicable ,other=other, SC=science(Art), SS=social science (sadist kilo),Tec=Technology N .and S.,med= medical(black lion)}
19	Category	Which program the client registerd for learning or not registed student	{Comm=community , EXSTU=extension student,other=other, PART= parent of staff,RESTU= regular student.STAFF=staff, SUSTU= summer student}
20	**HIV**	Laboratory result of HIV status	{neg=Negative , pos=Positive}

Appendix 3: Summary Of Confusion Matrix used for Experimentation

1) J48 algorithms with 10- fold cross validation with all attribute

a b <-- classified as

7040 636 | a = Neg

54 7666 | b = Pos

2) Summary of pruned J48 algorithms with minNumObj = 200 with 10 fold cross validation

a b <-- classified as

6033 1643 | a = Neg

1797 5923 | b = Pos

3) Result of re-evaluation of best J48 decision tree with 3000 dataset

=== Re-evaluation on test set ==User supplied test set Relation: finalt estcaseI nstances: 3000

Attributes: 20

a b <-- classified as

71 8 | a = Neg

71 1388 | b = Pos

4) **The snap shot of Confusion matrix Of PART algorithm with 10-fold cross validation and** all **attribute**

a b <-- classified as

7182 494 | a = Neg

16 7704 | b = Pos

5) The confusion matrix of PART Algorithms with feature selected (9 attribute)

a b <-- classified as

6501 1175 | a = Neg

795 6925 | b = Pos

6) The experimental result of SMO algorithms confusion matrix

a b <-- classified as

1903 437 | a = Neg

458 21 | b = Pos

a) 70-30 percenatge splite

a b <-- classified as

6283 1393 | a = Neg

1593 6127 | b = Pos

b) 10-fold cross validation

Appendix 4. Sample J48 Decision Tree With 10 Fold Cross Validation

J48 pruned tree

```
PrevTest =  yes-ve
|  EverHSex = Yes
|  |  Sex = M
|  |  |  Age = A4
|  |  |  |  MarStatus  = Nmar
|  |  |  |  |  ReaHere = plan
|  |  |  |  |  |  Residance = R14
|  |  |  |  |  |  |  UseCond = Ne: Neg (118.0)
|  |  |  |  |  |  |  UseCond = Na
|  |  |  |  |  |  |  |  Year = Y3: Neg (120.0)
|  |  |  |  |  |  |  |  Year = Y1: Neg (7.0)
|  |  |  |  |  |  |  |  Year = Y4
|  |  |  |  |  |  |  |  |  ExpoTime = dnotkn: Pos (11.0/2.0)
|  |  |  |  |  |  |  |  |  ExpoTime = less1m: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = noexpo: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = 4to6m: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = less1mto3m: Pos (0.0)
|  |  |  |  |  |  |  |  |  ExpoTime = greater6m: Neg (2.0)
|  |  |  |  |  |  |  |  Year = NY: Neg (17.0)
|  |  |  |  |  |  |  |  Year = Y2: Neg (4.0)
|  |  |  |  |  |  |  |  Year = Y6: Neg (2.0)
|  |  |  |  |  |  |  |  Year = Y5: Neg (1.0)
|  |  |  |  |  |  |  UseCond = So
|  |  |  |  |  |  |  |  HISTI = No
|  |  |  |  |  |  |  |  |  EduStatus = other
|  |  |  |  |  |  |  |  |  |  Category = RESTU: Neg (2.0)
```

Appedix 5: Sample PART Decision Rule List With 10 Fold Cross Validation

PART decision list

PrevTest = yes+ve AND

Steady = 1p AND

Category = RESTU AND

SessType = Indv AND

ReaHere = status AND

ExpoTime = greater6m: Pos (320.0/2.0)

EverHSex = No AND

MarStatus = divor AND

ReaHere = causal: Pos (63.0/1.0)

EverHSex = No AND

MarStatus = divor AND

PrevTest = yes+ve: Pos (46.0)

EverHSex = No AND

MarStatus = wido AND

PrevTest = No: Pos (30.0)

EverHSex = No AND

PrevTest = yes-ve AND

ReaHere = status: Neg (103.0)

PrevTest = yes+ve AND

Category = Comm AND

EverHSex = Yes: Pos (121.0)

PrevTest = yes+ve AND

Steady = 1p AND

Category = 99.0: Pos (291.0/2.0)

EverHSex = No AND

ReaHere = status AND

Year = Y3 AND

ExpoTime = dnotkn: Pos (62.0)

EverHSex = No AND

ReaHere = plan AND

MarStatus = Nmar AND

Residance = R4: Neg (239.0)

EverHSex = No AND

ReaHere = plan AND

MarStatus = Nmar AND

UseCond = Ne AND

ExpoTime = noexpo: Neg (235.0)

EverHSex = No AND

UseCond = Na AND

MarStatus = Nmar AND

EduStatus = Sesch AND

Steady = 1p AND

ReaHere = plan AND

ExpoTime = dnotkn: Neg (169.0)

ReaHere = causal AND

Year = Y3 AND

PrevTest = No AND

EverHSex = Yes AND

Category = RESTU AND

LastSex = Y AND

Age = A4: Pos (75.0)

ReaHere = causal AND

Year = Y3 AND

PrevTest = No AND

EverHSex = Yes AND

UseCond = Na AND

MarStatus = wido: Pos (55.0)

ReaHere = causal AND

Year = Y3 AND

PrevTest = No AND

EverHSex = Yes AND

UseCond = Ne AND

HISTI = No AND

EduStatus = Illitate: Pos (113.0)

Appendix 6: sample selected interesting rule discussed as findings

Rule 1: EverHSex = No AND ReaHere = status AND Year = Y3 AND ExpoTime = dnotkn: Pos (62.0)

Rule 2: ReaHere = causal AND Year = Y3 AND PrevTest = No AND EverHSex = Yes AND Category = RESTU AND LastSex = Y AND Age = A4: Pos (75.0)

Rule 3: ReaHere = causal AND Year = Y3 AND PrevTest = No AND EverHSex = Yes AND UseCond = Na AND MarStatus = wido: Pos (55.0)

Rule 4: ReaHere = causal AND Year = Y3 AND UseCond = Ne AND PrevTest = No AND HISTI = No AND Age = A6: Pos (111.0/3.0)

Rule 5: ReaHere = risk AND LastSex = N AND Year = Y3 AND Residance = R14 AND Sex = F AND EverHSex = Yes AND Age = A6 AND

Rule 6: PrevTest = No AND Age = A6 AND Year = Y3 AND Category = Other AND Sex = M: Pos (106.0/6.0)

Rule 7: PrevTest = yes-ve AND Sex = M AND Age = A3 AND HISTI = NA AND EduStatus = ugrad: Neg (75.0)

Rule 8: Sex = F AND EverHSex = Yes AND Year = Y1 AND UseCond = Ne AND EduStatus = ugrad: Pos (10.0/2.0)

Rule 9: Sex = F AND EverHSex = No AND Residance = R1 AND PrevTest = No: Pos (23.0)

Rule 10: PrevTest = No AND Residance = R14 AND Year = Y3 AND HISTI = No AND UseCond = So AND Age = A3 AND ExpoTime = less1m: Pos (12.0)

Rule 11: Residance = R14 AND HISTI = No AND UseCond = So AND Employed = Yes AND Age = A3 AND ExpoTime = greater6m: Pos (12.0)

Rule 12 :PrevTest = No AND Residance = R14 AND HISTI = No AND ExpoTime = 1mto3m AND Age = A4: Neg (15.0)

Rule 13 : PrevTest = No AND Residance = R14 AND HISTI = No AND LastSex = N AND EduStatus = Sesch AND Employed = NO AND Category = RESTU: Neg (20.0)