ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Applicability of Data Mining Techniques to Support Voluntary Counseling and Testing (VCT) for HIV: The Case of Center for Disease Control and Prevention (CDC)

BY
Biru Asmare

January 2009

A Thesis Submitted To The Graduate Studies Of Addis Ababa University In Partial Fulfillment Of The Requirements For The Degree Of Masters Of Science In Information Science.

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

Applicability of Data Mining Techniques to Support Voluntary Counseling and Testing (VCT) for HIV: **The Case of Center for Disease Control and Prevention (CDC)**

BY
BIRU ASMARE

Name and Signature of Members of the Examining Board

_____

_____

_____

Chairman,Faculty Signature Date

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First and foremost, I thank God for he is always besides me during my pleasure and trouble times. I am also deeply grateful and thankful to my advisor Ato Ermias Abebe for his guidance, encouragement and constructive suggestions throughout this thesis work.

Secondly and foremost I would like to acknowledge CDC, OSSA, and Addis Ababa HAPCO staff. Special thanks to W/o Yewubnesh Hailu, and Tamiru Ayana, besides their effort in providing necessary information in the domain area, made things easy in getting access to the VCT Epi-Info database.

I would also like to acknowledge all my classmates, and friends for all constructive comments and suggestions in the overall thesis work. I have appreciation and thanks to Queens' college management and staff for their support and patience.

Finally, special thanks go to my wife w/o Tigist Mekonnen for her devotion in handling all family related issues on behalf of me for the past two years.

ABBREVIATIONS

AIDS	Acquired Immunodeficiency Syndrome
AAHAPCO	Addis Ababa HIV/AIDS Prevention and Control Office
ANN	Artificial Neural Network
ART	Anti Retroviral Treatment
ARC	AIDS Resource Center
ARFF	Attribute Relation File Format
CDC	Center for Disease control and prevention
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma separated Value
CITC	Client-initiated testing and counseling
DM	Data Mining
EM	Expectation Maximization
HAPCO	HIV/AIDS Prevention and Control Office
HIV	Human Immunodeficiency Virus
ICT	Information and Communication Technology
KDD	Knowledge Discovery in Databases
MARP	Most At Risk Populations
MOH	Ministry Of Health
MTCT	Mother to Child Transmission
NGO	Non Governmental Organization
OLAP	Online Analytic Processing
OSSA	Organization for Social Services for AIDS

PSI	Population Services International
PITC	Provider-initiated testing and counseling
SVM	Support Vector machine
UNAIDS	Joint United Nation Program on HIV/AIDS
VCT	Voluntary Counseling and Testing
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Data mining is emerging as an important tool in many areas of research and industry. Companies and organizations are increasingly interested in applying data mining tools to increase the value added by their data collections systems. Nowhere is this potential more important than in the healthcare industry. As medical records systems become more standardized and commonplace, data quantity increases with much of it going unanalyzed. Data mining can begin to leverage some of this data into tools that help health organizations to organize data and make decisions.

Data related to HIV/AIDS are available in VCT centers. A major objective of this thesis is to evaluate the potential applicability of data mining techniques in VCT, with the aim of developing a model that could help make informed decisions. Using the dataset collected from OSSA, which is supported by CDC, and CRISP-DM as a knowledge discovery process model findings of the research are presented using graphs and tabular formats

For the clustering task the K-means and EM algorithms were tested using WEKA. Cluster generated by EM were appropriate for the problem at hand in generating similar group. According to the results of these experiments it was possible to see similar groups from VCT clients. The gender, martial status, and HIV test result, and education has shown patterns.

For the classification task, decision tree (J48 and Random tree) and neural network (ANN) classifier are evaluated .Although ANN shows better accuracy than decision tree classifier, the decision tree (J48) is appropriate for the dataset at hand and is used to build the classification model. Finally, cluster-derived classification models are tested for their cross-validation accuracy and compared with non cluster generated classification model.

The outcomes of this research will serve users in the domain area, decision makers and planners of HIV intervention program like CDC and MOH.

CHAPTER ONE
INTRODUCTION

1.1 Back ground

Large amounts of data have been collected routinely in day to day activities of business, administration, banking, social and health services, environmental protection, and security. There is an explosive growth in generating data in all fields. The amount of data stored in the world's database is expected to grow by double every twenty months (Witten and Frank, 2000).

Even if we are such huge amount of data it is not possible to say the data do not decrease the striving for knowledge. This is because the processing power of evaluating and analyzing the data did not follow this massive growth. As a result of this phenomenon, a tremendous volume of data is still kept without being studied.

Data mining, a research field that tries to ease this problem, proposes some solutions for the extraction of significant and potentially useful patterns from these large collections of data.

In earlier time the traditional method of turning data into knowledge relies on manual analysis and interpretation (Fayyad et al., 1996).It was also possible to identify and utilize information hidden in data via query generators and data interpretation systems. In both

cases the quality of the extracted information depends on the user's interpretation of the results and is thus vulnerable for errors from subjectivity.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction (Two crows, 2005).

Recent advance in ICT have made data easy to use and cheap to store and exchange. Databases across the world contain data that exist in digitized text document, video and audio files, and financial transactions. Different public and Private organization gather, process, disseminate and store data using ICT (Corcoran and Ware, 2001).

In the last 10 years many data mining systems and products have been invented so vast that they can be used in many applications; examples include predicting costs of corporate expense claims, in risk management, in financial analysis, in insurance, in process control, in manufacturing, in healthcare, and in other fields (Han and Kamber, 2001).

The Healthcare industry, which is among the most information intensive areas, is one of the focus points in data mining research. Medical data keep growing on a daily basis. The ability to use these data to extract useful information for quality healthcare is crucial.

One of the challenging problems in relation to health in the world is Human Immunodeficiency Virus (HIV) /Acquired Immune Deficiency Syndrome (AIDS).

HIV/ AIDS is a complicated issue and a very hot global agenda. The number of people living with HIV has risen from around 8 million in 1990 to more than 33 million, and is still growing. Around 69% of people living with HIV are in sub-Saharan Africa in the year 2007 (UNAIDS, 2007).

According to the report of UNAIDS (2007), around two and a half million adults and children became infected with HIV, the virus that causes AIDS. The year also saw more than two million deaths from AIDS, despite recent improvements in access to antiretroviral treatment.

Ethiopia is one of sub-Saharan Africa country also affected by HIV. The first case of HIV/AIDS in Ethiopia was diagnosed in 1986. According to the MOH, approximately 3.2 million Ethiopians are living with HIV/AIDS, with an adult prevalence of 7.7% in urban and 0.9 % in rural. The U.S. Census Bureau estimates that life expectancy in Ethiopia will decline to about 42 years due to AIDS by 2010; without AIDS, life expectancy would be 55 years.

The latest studies show that the epidemic is concentrated in urban areas, transport corridors, and most at risk populations (MARPs) (Center for Disease control and prevention, 2007). The HIV/AIDS Policy was formulated by the Ministry of Health (MOH) and adopted by the Council of Ministers in 1998 (Wubit, 2007)

Addis Ababa, one of the cities in Africa, is highly affected by the HIV/AIDS pandemic with an adult prevalence rate of 7.5 %. The overall challenge of curbing the epidemic in the city is the need for a coordinated, complementary and harmonious joint effort of all stakeholders in Addis Ababa (UNAIDS, 2007).

In response to the HIV/AIDS epidemic, it was decided at a national level to establish HIV/AIDS Prevention and Control Offices (HAPCO's) in every region of Ethiopia. As a result of this, Addis Ababa HAPCO (AAHAPCO) was founded in January 2001 to coordinate all HIV/AIDS prevention and control programs in the Addis Ababa region.

All stakeholders in Addis Ababa have introduced various initiatives to respond to the HIV / AIDS epidemic and have scaled up all efforts on HIV and AIDS prevention, care and support, including the provision of antiretroviral treatment. One of such program is Voluntary Counseling and Testing for HIV (VCT).

Most VCT centers in Ethiopia are supported by CDC. The OSSA VCT center is one of the center that is supported by CDC .The center provides counseling services and collect records using the Epi-Info database. Epi Info is a database which is used by epidemiologists and other public health and medical professionals to develop Epidemiologic statistics, tables, graphs, and maps are produced with simple commands. The record is used to generate reports to respective bodies.

For health related issues it is important to be able to interpret patient information and categorize similar people together. In doing this the data mining techniques play an important role. Because the methodology in data mining can help to extracts hidden predictive information from large databases like Epi_Info.

Data mining models are based on one of two kinds of learning, supervised and unsupervised. Supervised learning functions are typically used to predict a value. Unsupervised learning functions are typically used to find the intrinsic structure, relations, or affinities in a body of data but no classes or labels are assigned a priori (Han and Kamber, 2001).

Clustering, unsupervised learning is a technique useful for exploring data. It is particularly useful where there are many cases and no obvious natural groupings. Here, clustering data mining algorithms can be used to find whatever natural groupings may exist.

In a classification problem, you have a number of cases and wish to predict which of several classes each case belongs to. Each case consists of multiple attributes, each of which takes on one of several possible values in each case. All but one attribute is a predictor attribute, and one is the target attribute. Each of the target attribute's possible values is a class to be predicted on the basis of that case's predictor attribute values.

1.2 Statement of the problem

According to AAHAPCO (2007), the HIV/AIDS epidemic is claiming the lives of the most productive, energetic and educated segments of the population in the city. The adult prevalence rate of HIV/AIDS in the city has increased from 7.2% in 2004 to 7.5% in 2007. It is projected to be 7.9%, 8.5% and 9.2 % in the years 2008, 2009, and 2010 respectively. Effort is being made by governmental and non governmental organizations to reduce the vulnerability of adults to HIV/AIDS and to minimize this increasing rate through counseling.

VCT centers collect huge information about clients which could be used for counseling. However, they do not use the accumulated data for the counseling service they provide. They just send this huge information to the CDC as there are no mechanisms that support them in extracting hidden knowledge relevant for their counseling tasks. CDC then enters the data collected from different VCT centers to its central database. This database requires authentication from USA to be accessed. Through such authentication, CDC uses the database to generate statistical reports on HIV/AIDS for different bodies.

The underlying research problem that necessitated this research is the fact that, although huge amount of electronic data on HIV/AIDS is available at VCT centers, they are not using it in a way that supports their objectives. Decision making bodies on HIV/AIDS in the country are also not wisely using this data for making informed decisions. Thus, the huge data remains unutilized to solve the problems faced by the society due to lack of research in deploying appropriate data analysis and mining tools.

Only one attempt has been made by Abraham (Abreham, 2005) to apply data mining application on HIV/AIDS data collected at VCT centers for the purpose of identifying determinant risk factors of HIV and to find their association rule. From the different kinds of data mining functionalities (characterization, association, classification, clustering, outlier and trend analysis, etc), Abreham used association rule mining through Apriori algorithm. The data set for his experiment were 15801 records collected for two years (09/12/2002-09/12/2004) (Abreham, 2005).

Mining Association Rules is a two-step approach, frequent itemset generation and rule generation. But frequent itemset generation is computationally expensive process, because candidate generation can result in huge candidate sets and it will also result in multiple scan of database. For many frequent-itemset algorithms, main memory is the critical resource. Since discovering frequent itemsets requires a lot of computation power, memory and I/O the selected method lacks scalability.

In assessing the applicability of data mining techniques on VCT dataset, Abreham recommended to use large dataset rather than a small percentage of the two years VCT data and to include and study variables that are not considered as important by domain experts (Abreham, 2005).

VCT centers have now five-year data, which is larger than the data used in Abrham's research. Mining such large data requires algorithms other than Association Rule Mining.

The work of Shegaw (2002) is another attempt in the applicability of data mining on health data. This work focuses on predicting child mortality and is not related to HIV/AIDS counseling. In addition, previous researches in health were conducted from very small portion of databases using simple statistical techniques as a data analysis tool (Shegaw, 2002). As a result, decision making process made by all stakeholders in HIV/AIDS issue are not supported by sound tools and techniques that could enable them to extract hidden knowledge and patterns from records (Abreham, 2005).

The lack of adequate research in the applicability of Data Mining techniques to support HIV/AIDS counseling justifies a new research that can handle the large data available at VCT centers. This research may fill the gaps indicated above around the health sector (particularly in HIV/AIDS) in using data mining techniques.

The research question of this study is, from the very large data available in health databases like VCT client data records, how we can extract the hidden information which has significant strategic importance for HIV/AIDS intervention programs. Today, organizations including VCT centers are dealing with large databases. Extracting useful information from this large data is a big challenge. “When there are millions of trees, how can one draw meaningful conclusion about the forest?” (Two crows, 2005).

This research tries to address the following question:

1. What hidden knowledge is there in the large data available at VCT centers?
2. What types of Clients are residing in the VCT database?

3. How data mining techniques can be applied in supporting VCT centers and other bodies working to prevent and control HIV/AIDS?
4. Will the identification of different clusters of VCT clients enable VCT centers and stakeholders to tailor their Counseling strategies
5. Will it be possible to automatically predict the class label of instances of VCT client?

1.3 Objective of the study

The general and specific objectives of the proposed study are the following:

1.3.1 General Objectives

The general objective of the research is to investigate the potential applicability of data mining techniques on VCT client dataset in developing a model that could support VCT in identifying patterns , there by enabling policy makers, Health officers, Donors, NGO's, and Counselors to make informed decision in the effort to plan HIV/AIDS intervention programs.

1.3.2 Specific objectives

In order to achieve the general objective indicated above, the research has the following specific objectives.

- Conduct a thorough review of literature on the existing data mining technologies and methods in general and their application in the health sector.

- Select the data mining tools and techniques to be used based on the type of data mining functions to be performed.
- Identify different data mining application tools supporting clustering and classification and which are more appropriate to the problem domain, and select the best tool.
- Select and extract the data set required for analysis from VCT centers in Addis Ababa
- Prepare the data, for analysis and model building, by cleaning, extracting and transforming the data into a format suitable for the selected data mining algorithm.
- Apply clustering algorithm to group VCT clients.
- Apply the selected data mining classification algorithms to build and train the classification models that classify instances of VCT clients into one of the class labels identified by the clustering algorithm.
- Test and compare the resulting performances of the clustering and classification models
- Build and test the model using the selected tools and techniques.
- Report results and make recommendations for further research.

1.4 Research Methodology

A number of knowledge discovery process models are used in data mining research. The CRISP-DM, and Fayyad et al, are one of basic approach followed in KDD (Olson and Delen, 2008). CRISP-DM is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study. CRISP-DM was proposed in the mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining.

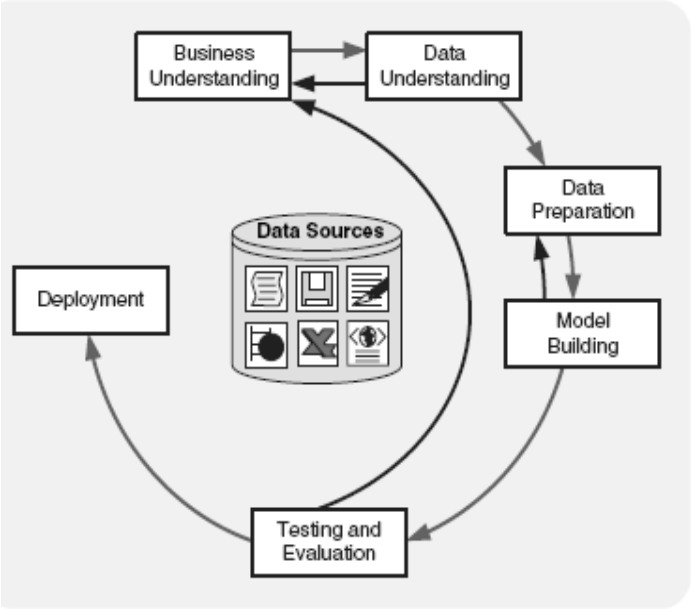


Figure 1: Phase of CRISP-DM (Olson and Delen, 2008)

As it is indicated in Figure 1, the Phases in CRISP-DM are Business understanding, Data understanding, Preparation, Modeling and Evaluation. A more detail about the methodology used is covered under the research methodology chapter of this research.

1.4.1 Literature Review

For a better understanding of a given problem, reviewing the relevant literatures on data mining concepts, techniques and applications are carried out. The application of data mining in the area of health is deeply analyzed and assessed. Various books, journals, magazines, articles, and papers from different sources including the World Wide Web are consulted in order to have understanding the concepts and techniques of data mining and KDD. These sources provide sound theoretical and practical background to the research.

1.4.2 Fact Finding

In order to define, analyze, and understand the research problem clearly and properly, data was collected by observation and interviewing concerned experts in counseling. This helps to have information about the views and opinions of domain experts on VCT dataset and to select features with domain experts. In addition observation was conducted to understand the counseling process.

1.4.3 Identification and Selection of Target Dataset

The main issue at this point was the identification and collection of the data sets that are relevant to the problem under question. The VCT records' database is an appropriate data source for the research. The dataset was collected from Organization for social Services for AIDS (OSSA).The raw data to be used for this research work is basically on the information obtained from HIV counseling and testing records. The record has a pre-test and post-test counseling session that collects data like martial status, educational experience, occupation title, how hear about VCT, sex, age, and test result.

1.4.4 Data Preparation

In this stage, collected data was arranged into a form that will be suitable for the particular data-mining software. To have a normalized data the necessary data preprocessing activities like handling inconsistent and missing values, noisy data,

deriving new fields and summarization was performed. This helps to effectively and efficiently apply data mining tools and algorithms on the target data.

1.4.5 Model Building/Data Analysis

One of the major tasks at this stage is the selection of appropriate data mining tool and modeling the collected data using appropriate data mining techniques. After the selection of the appropriate clustering and classification data mining tools and algorithms as well as completing the cleaning, formatting and transformation of the data, different models was built through the training and testing datasets. Once the models are built and trained, evaluating and interpreting the results of the selected model was done. Using the confusion matrix and the suggestion of domain experts the best model was selected.

1.4. Scope and Limitation of the Study

The scope of the research is to appraise the potential applicability of data mining in supporting HIV/AIDS intervention program for VCT centers in Addis Ababa that are supported by CDC. While the findings of this research work can fairly be considered and adopted as relevant to implement the potential applicability of data mining in other sector of health giving similar services related to HIV/AIDS, the scope of the current experimental research is applicable to only to the cases OSSA in Addis Ababa. The research explore the applicability of the clustering and classification function of data mining to VCT pre-test and post test counseling dataset. In addition this research does not include the case of antiretroviral treatment cases and the different types of HIV virus.

The major limitations while undertaking this research was on the data set. Data for this research are only from VCT centers. The dataset do not includes records obtained from mobile testing centers. Moreover lack of relevant literature on application of data mining on VCT client data and little prior experience of the researcher to different kind of data mining techniques and tools is another limitation.

1.5 Application of Research Findings

Although this research is primarily initiated to fulfill academic requirements, the result obtained from this study could be applied in various areas.

Firstly, it will provide additional dimension of research for organizations like CDC, MOH, AHAPCO, and OSSA who are involved in research and planning.

Secondly, the result of the research could be used by policy makers who are expected to make decision about HIV/AIDS prevention intervention program and set policies based on research findings

Thirdly, researchers who are involved on HIV/AIDS studies can use the output and recommendation to fill the gap.

The results from this research are expected to contribute a lot for HIV/AIDS intervention programs so that citizens will get quality of life and reasonable life expectancy by minimizing the risks that are caused by the virus. The research will also indicate the

applicability of data mining in health industry. Besides, it will initiate other researchers to further undertake research on similar problem domain.

1.6. Organization of the Thesis

The thesis is organized in to five chapters. These are introduction, the literatures review, research methodology, experimentation and discussion of Results, and conclusion & recommendation. Chapter one is an introductory chapter, which states the background of VCT, HIV/AIDS and data mining the research in need, a problem statement with research questions, research objectives, and research methodology.

Chapter two provides overview about data mining, models, algorithms and tools. In addition this chapter describe about HIV/AIDS and VCT. Literatures and similar research works in the application of data mining to health specifically to HIV/AIDS and VCT are discussed.

In chapter three details of methodology followed by the researcher is discussed. In the chapter four, the experimentation and discussion of results is presented and the last chapter addresses the conclusion and recommendation part based on the results of the experimentation. Finally references used by the researcher during the research work were included.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Data Mining

2.1.1 Introduction

In the information society, advance in ICT have made data easy to use and cheap to store and exchange. Database across the world contain data that exists in digitized text document, video and audio files, and financial transactions (Corcoran and Ware, 2001). Even if we are surrounded by huge data, we are facing difficulty in getting information and/or knowledge. The available data, in itself, is not enough for improving the work and to have meaning and relationship. We need to be able to transform the raw data into information which is useful for taking important business decisions.

Early pioneers , such as U. Fayyad, H. Mannila, G. Piatetsky-Shapiro, G. Djorgovski, W. Frawley, P. Smith, and others recognized the need of a new field which solve the problem of huge data with no value and lack of mechanisms to efficiently and effectively extracting information and knowledge from them (Cios et,al. ,2007).

As a field of study, scholars are explaining theories and definition of the field. They have come up with numerous definition of data mining. Some of them will be summarized as follows.

Witten & Frank (2005) defined data mining as *'Data mining is the extraction of implicit, previously unknown, and potentially useful information from data'*.

According to Pyle (2003) data mining is not only a tool work in which computer software will do and support giving. For him, "*data mining is a structured way of playing with data, of finding out what potential information it contains and how it applies to solving the business problem*."

As it is shown in Figure 2.1 below, Data mining field was introduced relatively recently in the 1990s and its history is rooted back to three family lines called classical statistics, artificial intelligence, or AI, and machine learning, which is more accurately described as the union of statistics and AI (Buchanan, 2006).

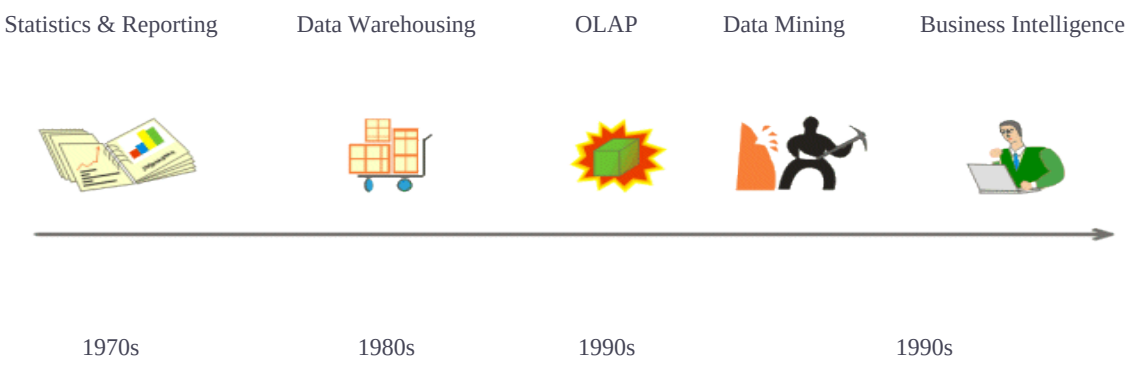


Figure 2. 1: Evolution of Data Mining Technology

(Business Intelligence to Marketing and Management (BI2M) service)

2.1.2 Why Data mining

As suggested by Cios et al. (2007) even if data mining came into existence in response to technological advances in many diverse disciplines, it is not an “umbrella” term coined for the purpose of making sense of data. Because other fields like statistics also deals with data. Cios et al. (2007) further argued that data mining is is a data driven approach, as opposed to model driven. In statistics, researchers frequently deal with the problem of finding the smallest data size that gives sufficiently confident estimates. Data mining deals with the opposite problem, namely, data size is large and we are interested in building a data model that is small but still describes the data well.

When we see the evolution of data mining, it does not mean that data mining is completely different from other disciplines and field of studies like statistics, Data warehousing, OLAP, and Business intelligence.

In relation to Statistics and Data mining, Witten and Frank (2000) cited in Tibebe (2005) have raised points on difference and similarity. DM usually deals with the data collected for other purposes where as Statistics deals with data collected using efficient strategies to answer specific question. Though, data mining techniques are better in the understanding of data analysis and utilization, it does not mean that data mining has replaced statistics. Rather, data mining is an extension of statistical methods.

Data mining takes advantage of advances in the fields of artificial intelligence. Both disciplines have been working on problems of pattern recognition and classification (Two Crows, 1999).

For organizational learning to take place, data from many sources must be gathered together and organized in a consistent and useful way hence, Data Warehousing. Many organizations use this information to support decision making activities (Han and Kamber, 2001).

Even if Data warehouse is not a requirement of data mining, it provides clean and integrated data for fruitful mining. Due to this there is some real benefit if the data is already part of a data warehouse. But, in this work the VCT data are not in a warehouse. It is better to use data mining because it can provide powerful tools (e.g., association, classification, clustering, and trend analysis) for analysis of data stored in data warehouses (Two crows, 2005).

OLAP, part of the spectrum of decision support tools, is used to answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. According to Two Crows (1999), OLAP Provides a very good view of what is happening, but can not predict what will happen in the future or why it is happening.

In this investigation Data mining is used because it is different from OLAP in that rather than verifying hypothetical patterns, it uses the data itself to uncover such patterns. More over we can use data mining in forecasting what may happen in the future, classifying

people or things into groups by recognizing patterns, clustering people or things into groups based on their attributes, associating what events are likely to occur together, and sequencing what events are likely to lead to later events.

2.1.3 Data Mining and Knowledge Discovery

Knowledge Discovery in Databases (KDD) is defined as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns of data (Fayyad, 1996). One of the crucial steps in KDD is Data-mining. KDD and Data-mining, however, are often used as synonyms.

For Benoit (2002) as it is quoted in Ayre (2006), KDD involves searching large databases, but it distinguishes itself from database querying in that it seeks implicit patterns in the data rather than simply extracting selections from the database.

As it is indicated in Figure 2.2 Data Mining, a step in the KDD process is a means to find patterns in large amounts of data by fitting models that are not necessarily statistical models. Traditional techniques may be unsuitable due to enormity of data, high dimensionality of data, and heterogeneous and distributed nature of data.

The knowledge discovery process is iterative, and involves steps with many decisions made by the user at each step.

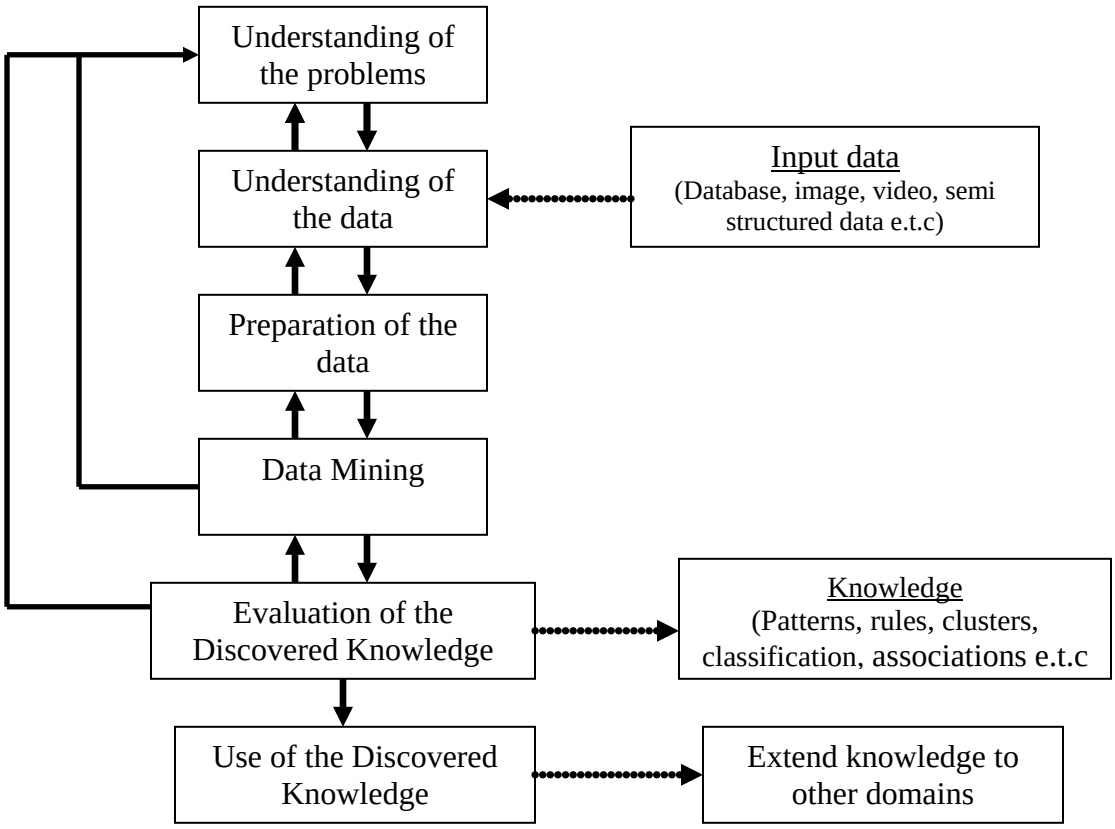


Figure 2. 2: Data Mining in Knowledge discovery process model (Cios et al., 2007)

2.2 Data Mining Models and Methods

Data mining could be applicable to any kind of information repository. This includes relational database systems, data warehouses, transactional databases, flat files, and the World Wide Web. The challenges and techniques of mining may differ for each of the repository systems (Han and Kamber, 2001).

According to Fayyad,U.et al.(1996) cited in Denkew (2003) Prediction and description are considered the two primary goals of Data mining in practice.

2.2.1 Models

Predictive modeling refers to the process of building a model that will permits the value of one variable to be predicted from the known value of other variables (Hand, 2005). The main goal of Predictive modeling is to predict the value of one column based on the value of other column. It can be used to forecast explicit values, based on patterns determined from known results. It is referred as supervised learning in which its functions are typically used to predict a value (Han and Kamber, 2001).Classification and regression are the most common Predictive modeling tasks.

In this investigation one of the known predictive task, Classification is going to be implemented on VCT data to develop a possible model that could support VCT.

Another modeling which deals with describing all of the data and discovering patterns and segments of the data is descriptive modeling (Hand, 2005).It is referred us Unsupervised learning, in which its functions are typically used to find the intrinsic structure, relations, or affinities in a body of data but no classes or labels are assigned a priori (Han and Kamber, 2001). Clustering and association rule are the known descriptive modeling task.

2.2.2 Data Mining Methods

In order to achieve the higher level goals of data mining, prediction or description, different data mining methods or tasks are suggested by scholars. Han and Kamber (2001) described the data mining tasks as concept/Class description, characterization and

description, association analysis, classification and prediction, cluster analysis, outlier analysis, and evaluation analysis.

Fayyad et al. (1996) listed as methods that include: classification, regression, clustering, summarization, dependency modeling, and change and deviation detection.

Literatures describe the data mining tasks/methods in a slightly different ways and variations exist on the basic approaches described below, some of the basic methods will be presented as described Kantardzic (2003).

Classification: discovery of a predictive learning function that classifies a data item into one of several predefined classes.

Regression: discovery of a predictive learning function, which maps a data item to a real-value prediction variable.

Clustering: is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.

Summarization: an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.

Dependency Modeling: finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.

Change and Deviation Detection: discovering the most significant changes in the data set.

2.3 Data Mining Algorithms and Tools

Data mining is the analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand et.al, 2001).

In Section 2.2.1 the two high-level primary goals of data mining in practice (prediction and description) were discussed. In the following sections data miming algorithms that are used to achieve goals will be reviewed.

2.3.1 Clustering Algorithms

Clustering is an active area of data mining research, in which numerous research groups are working on it by focusing on creating technique which helps to automatically group data into related clusters (Han and Kamber, 2001), (Berkhin, 2002).

A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words, members of a cluster are more like each other than they are like members of a different cluster (Fayyad, et al., 1996).

Clustering models are different from predictive models in that the outcome of the process is not guided by a known result, that is, there is no target attribute. Predictive models predict values for a target attribute, and an error rate between the target and predicted

values can be calculated to guide model building. With clustering models, the data density itself drives the process to a final solution, that is, determine clusters.

For Han and Kamber (2001), there are different kinds of techniques for creating clusters. These include partitioning (often using the k-means algorithm) and hierarchical methods (which group objects into a tree of clusters), as well as gridbased (methods that quantize the object space into a finite number of cells that form a grid structure), model based (methods that hypothesize a model for each of the clusters and find the best fit of the data to the given model), and density-based methods which are based on the notion of density and are used to discovering clusters of arbitrary shapes.

Paquet (2004) and Han and Kamber (2001) argued that, these clustering algorithms are based on the distance measure between two objects that are represented by a set of attribute-value pairs. The methods for computing the similarity or dissimilarity between two objects, to a large extent, depend on the nature of the attributes themselves and the characteristics of the objects that we need to model by the similarity function.

2.3.1.1 The K-Means Algorithm

The k-means algorithm (Hartigan 1975; Hartigan & Wong 1979) quoted in Berkhin (2002), is by far the most popular clustering tool used in scientific and industrial applications.

The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, *k*. Then the algorithm iterates till convergence:

Step 1: Specification o*f K.* In this step the algorithm place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2: Data Assignment, each data point is assigned to its *closest* centroid, with ties broken arbitrarily. This results in a partitioning of the data. This step is repeated, and each point is one again assigned to the cluster with the closest centroid.

Step 3: Relocation of "means", each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions and loop back to step 2.

As suggested by X. Wu et al. (2007) the K-Means algorithm has limitation. These are the selection of the initial K and its sensitivity to the presence of outliers, since "mean" is not a robust statistic. To avoid this problem effective preprocessing step is mandatory to remove outliers can be helpful.

Despite its drawbacks, k-means remains the most widely used partitional clustering algorithm in practice. The algorithm is simple, easily understandable and reasonably scalable.

2.3.1.2 Expectation Maximization Algorithms

According to Witten and Frank (2005) EM is a mixture based algorithm that attempts to maximize the likelihood of the model. EM models the distribution of instances probabilistically, so that an instance belongs to a group with a certain probability. EM does not actually calculate probabilities, instead it calculates densities. The assumption is made by EM that the attributes are independent random variables.

Foe Berkhin (2002) one important feature of probabilistic clustering is that mixture model can be naturally generalized to clustering heterogeneous data. This is important in practice, where an individual (data object) has multivariate static data (demographics) in combination with variable length dynamic data (customer profile).

EM can handle both numeric and nominal attributes. The first step, calculation of the cluster probabilities (which are the “expected” class values) is expectation; the second step is calculation of the distribution parameters, is “maximization” of the likelihood of the distributions given the data (StatSoft, 2004)

Han and Kamber (2001) argued that, unlike K-means Instead of choosing k in such a subjective manner, the expectation maximization (EM) algorithm can be used to automatically choose the number of clusters based on probability distribution estimation. The EM algorithm form what are known as mixture models – models that describe the data using statistical distributions. The steps of the algorithm are as follows:

1. Select an initial set of model parameters (randomly or otherwise)
2. Repeat: Iteratively refine the parameters (or clusters) based on the following two steps:
3. Expectation Step: For each object, calculate the probability that each objects belongs to each distribution.
4. Maximization Step: Given the probabilities from the expectation step, find the new estimates of the parameters that maximize the expected likelihood.
5. Until: The parameters do not change (or change below some threshold)

The following probability density function for a normal distribution is used to compute the cluster probability for each instance. In the case of a single independent variable with mean μ and standard deviation σ, the formula is:

$$F(x) = \left(\frac{1}{\left(\sqrt{2\pi\sigma}\right) e^{\frac{-(x-\mu)^2}{2\sigma^2}}} \right)$$

In the two-cluster case, we will have the two probability distribution formulas each having differing mean and standard deviation values.

EM algorithm performs maximum likelihood estimation for samples in mixture model. EM uses probability of cluster membership instead of a distance metric, and samples are not assigned to 1 cluster, but partially to different clusters (proportionally to distribution). EM is much more general than just “clustering”, it finds number of distributions generating data and builds “mixture models”.

2.3.1.3 Cluster Validity

For Cios et al, (2007) clustering is one of the unsupervised learning techniques, there should be a very careful way of assessing its results. Are the generated clusters reflective of the true nature of the data? This is a fundamental issue that permeates all clustering pursuits and profoundly impacts the practical usefulness of the technique.

For a cluster to be valid it should adhere to two fundamental requirements: Compactness, which expresses how close the elements in a cluster are and separability, the evaluation of how distinct the clusters are (Cios et al, 2007).

2.3.2 Classification Algorithms

.

Data classification according to Chen et al (1996) is the process which finds common properties among asset of objects in a database and classifies them into different classes, according to a classification model.

Benoit (Ayre, 2006) stated that Classification is composed of two steps: supervised learning of a training set of data to create a model, and then classifying the data according to the model. Some well-known classification algorithms include Bayesian Classification (based on Bayes Theorem), decision trees, neural networks, k-nearest neighbor classifiers, and genetic algorithms.

2.3.2.1 Decision Trees

As it is suggested by Berry & Linoff (2004), a decision tree is a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules.

For Kantardzic (2003) the decision-tree representation is the most widely used logic method. It is supervised learning methods that construct decision trees from a set of input-output samples. A typical decision-tree learning system adopts a top-down strategy that searches for a solution in a part of the search space.

As suggested by Cios et al (2007), decision tree consists of nodes and branches connecting the nodes. The nodes located at the bottom of the tree are called leaves and indicate classes. The top node in the tree, called the root, contains all training examples that are to be divided into classes. All nodes except the leaves are called decision nodes, since they specify decision to be performed at this node based on a single feature. Each decision node has a number of children nodes, equal to the number of values that a given feature assumes.

Tree Induction: For Gargano and Raggad (1999) decision trees are used to predict and/or classify. There are two phases, the training and implementation. During the training phase, the data set is partitioned iteratively. During each pass (i.e. iteration), the data set is split on that feature (or attribute) that produces the most effective classification. Only those factors most significant to the effective partitioning are used. The implementation phase then produces decision rules which are equivalent to the

partitions (or branching) created during the training phase. These rules are used to generate new information when presented with novel situations.

Decision tree attribute selection :Decision trees become incomprehensible when their size grows (Cios et al.,2007). Selecting the most discriminatory (significant) feature should be done to solve the problems. Attribute selection is normally done by searching the space of attribute subsets, and evaluating each one. This can be achieved by using a variety of techniques. Information gain attributes evaluation and gain ration attribute evaluations are some of the methods mentioned by Witten and Frank (2005).

At each node, available attributes are evaluated on the basis of separating the classes of the training examples. An evaluation (goodness) function is used. For this purpose as suggested by Han and Kamber (2001) information gain (ID3/C4.5), information gain ratio, and gini index (CART) are used.

Decision trees become incomprehensible when their size grows. To remedy this problem pruning techniques should be used to make decision trees more comprehensible (Cios et al., 2007).

As suggested by Han and Kamber (2001), to avoid the problem of overfitting decision trees are pruned down in such a way that there is no significant loss of classification accuracy. The pruning process can be handled by two ways, pre pruning and post pruning, depending on when the pruning occurs during the growth process of the tree.

In pre pruning, the growth of the tree stops when it is determined that no attribute will significantly increase the information gain in the process of classifying the data. While in post pruning, involves already-constructed trees. Often, the complexity of the tree is compared with the observed loss in classification accuracy in order to make a decision about how much of the tree branches should be eliminated.

2.3.2.2 Artificial Neural Network (ANN)

According to Kantardzic (2003), an artificial neural network is an abstract computational model of the human brain. It has the ability to learn from experiential knowledge expressed through inter unit connection strengths, and can make such knowledge available for use. The same author argued that, the use of artificial neural networks has the following capabilities.

A typical neural network is composed of a potentially large number of neurons arranged in three different conceptual layers: an input layer representing the input variables, one or more hidden layers, and an output layer representing the output variables.

A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response modes (Two Crows, 2005).

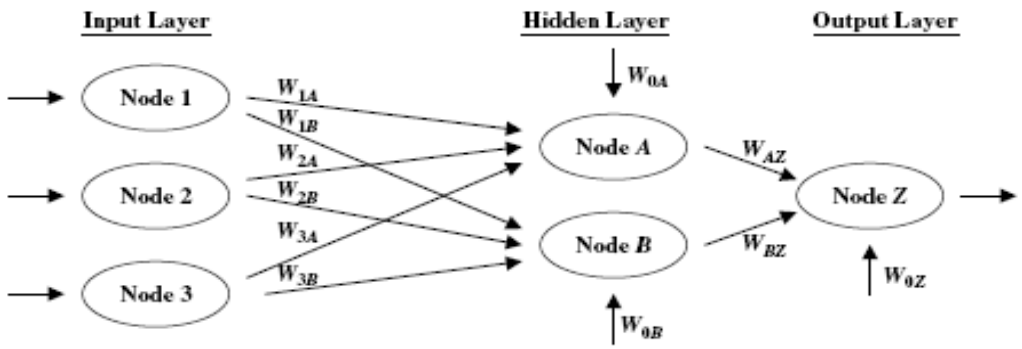


Figure 2. 3: A simple ANN with two hidden layers (Larose, 2005)

As it is indicated in Figure 2.3, W1A, W2B…W3B represents weights to the nodes. Every node in a given layer is connected to every node in the next layer, although not to other nodes in the same layer. Each connection between nodes has a weight (e.g., W1A) associated with it. At initialization, these weights are randomly assigned to values between zero and 1(Daniel, 2005).

Daniel (2005) further explained that the number of input nodes usually depends on the number and type of attributes in the data set. The numbers of hidden layers, and the number of nodes in each hidden layer, are both configured by the user.

On computation of the weight values and methods used, (Two Crows Corporation, 2005) described it as 'The connection weights (W's) are the unknown parameters which are estimated by a training method (back propagation). First a combination function (Usually summation) produces a linear combination of the inputs and the connection weights into a single scalar value. This will give the net for a given node.Once the net value for each node is known; it will be used as an input to the activation function (usually sigmoid function) which is used to generate the output signal from the weighted average of inputs. For the total output node Z, Net $_Z$ can be calculated.

The sigmoid function value of Net_z is out put value of the neural network for the first pass through the network and represents the value predicted for the target variable for the first observation**.**

Types of Learning: Neural networks have three main modes of operation – supervised, reinforced and unsupervised learning. In supervised learning the output from the neural network is compared with a set of targets, the error signal is used to update the weights in the neural network. Reinforced learning is similar to supervised learning however there are no targets given, the algorithm is given a grade of the ANN performance. Unsupervised learning updates the weights based on the input data only. The ANN learns to cluster different input patterns into different classes.

Neural network structures: There are 3 main types of ANN structures -single layer feed forward network, multi-layer feed forward network and recurrent networks. The most common type of single layer feed forward network is the perceptron. Other types of single layer networks are based on the perceptron model (Larose, 2005).

Inputs to the perceptron are individually weighted and then summed. The perceptron computes the output as a function F of the sum. The activation function, F is needed to introduce nonlinearities into the network. This makes multi-layer networks powerful in representing nonlinear functions.

Learning rate and momentum: The learning rate and the momentum are the constants chosen to help move the network weights towards a global minimum for sum squared error (SSE).

According to Larose (2005),the learning rate is a constant chosen to help us move the network weights toward a global minimum for SSE. However, what value should take? How large should the weight adjustments be?

In adjusting the learning rate, when it is very small, the weight adjustment tends to be very small. The network probably will take an unacceptably long time to converge. If the learning rate is large, then it will tend the network algorithm overshoot the optimal solution.

With respect to the momentum Larose (2005) explained that the back-propagation algorithm is made more powerful through the use of this term. Essentially, the momentum term represents inertia. Large values of the momentum will influence the adjustment in the current weight.

A momentum component will help to dampen the oscillations around optimality mentioned earlier, by encouraging the adjustments to stay in the same direction.

Advantage and Limitation: The main advantage of neural networks is that it is possible to train a neural network to perform a particular function by adjusting the values of connections (weights) between elements. For example, if we wanted to train a neuron

model to approximate a specific function, the weights that multiply each input signal will be updated until the output from the neuron is similar to the function.

Artificial neural networks (ANN) have memory. The memory in neural networks corresponds to the weights in the neurons. Neural networks can be trained offline and then transferred into a process where adaptive learning takes place.

The main disadvantage of ANN is they operate as black boxes. The rules of operation in neural networks are completely unknown. Another disadvantage is the amount of time taken to train networks. It can take considerable time to train an ANN for certain functions.

2.3.3 Data Mining Tools

As suggested by Pyle (2003), different good algorithms have been already developed in the past. For a researcher who needs to use the techniques of data mining to solve business problems, the algorithms can be used. These algorithms are taken out of the laboratory, wrapped in robust and reliable commercial packaging, tested for usability, and delivered with help screens, training manuals, tutorials, and instruction. These are nothing but tools.

The same author argued that data mining algorithms are the core mathematical and logical structures that direct and determine a specific computational approach to

examining data. Data mining tools are the commercially wrapped data mining algorithms ready for use in a business setting.

As shown in Two Crows (2005) choosing the right data mining product means finding a tool with good basic capabilities. The selected tool should have an interface that matches the skill level of the people who will be using it, and features relevant to your specific business problems.

According to Pyle (2003) data mining tools and software suites come in a variety forms. Some support multiple tools within integrated suites; others support single algorithms. The choice should be based on data mining functions and methodologies, data types, data sources, and scalability. This is because data mining tools perform data analysis to uncover important data patterns Han and Kamber (2001).

2.4 Challenges of Data mining

Data mining is a relatively new field and there are many challenges to be faced. Extracting useful information from data can be a complicated and sometimes a difficult process (Lee and Siau, 2001).

A system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and perform significant worse when a little noise is added to the training set. For Lee and Siau (2001), the most prominent problems and challenges of

data mining are noisy data, difficult training set, the nature of the database, and the size of the database.

2.5 Voluntary Counseling and Testing for HIV (VCT)

2.5.1 Overview of HIV /AIDS

For the past two decades, the world is facing a challenge by a virus which is causing a disease that killed millions of people and which looks likely to kill millions more. This virus is HIV. The HIV virus will damages the immune system, and causes a variety of symptoms known as AIDS after a period of time (UNAIDS, 2007).

According to AVERT (2007), the term epidemic is used when HIV/AIDS is widespread in a community. In 2007, advances in the methodology of estimations of HIV epidemics applied to an expanded range of country data have resulted in substantial changes in estimates of numbers of persons living with HIV worldwide.

In relation to the typology of HIV epidemics, for WHO and UNAIDS (2007) there are three levels of HIV epidemics Low level HIV epidemics if the HIV prevalence has not consistently exceeded 5% in any defined sub-population, Concentrated HIV epidemics If the HIV prevalence is consistently over 5% in at least one defined subpopulation but is below 1% in pregnant women in urban areas, and Generalized HIV epidemics if the HIV prevalence consistently over 1% in pregnant women.

2.5.2 HIV/AIDS in Ethiopia

Ethiopia is among the countries most heavily affected by the HIV/AIDS epidemic with the third largest population of HIV-infected persons living in Africa. The HIV, which is complicated issue and a very hot global agenda, was first detected in Ethiopia in 1984 and the first two AIDS cases were reported in 1986. A National HIV/AIDS taskforce was established in 1985 and the National AIDS Control Program (NACP) was established at a Department level at the MOH in 1987. HIV/AIDS surveillance activities began in 1989 (Wubit, 2007).

There are many factors that promote the spread of the disease including the presence of sexually transmitted infections, gender inequality, multiple sexual partners, prostitution, and men with disposable income, alcohol, unsafe blood transfusion, and transmission from infected mothers to their fetus/child during pregnancy and breastfeeding.

2.5.3 The Impact of HIV /AIDS

Every day, more than 6 800 people become infected with HIV and more than 5 700 die, mostly because they have no access to HIV prevention, treatment and care services. Despite progress made in scaling up the response over the last decade, the HIV pandemic remains the most serious infectious disease challenge to global public health(UNAIDS,2007).

The HIV/AIDS epidemic impacts on all spheres of life. One of the most significant features is its concentration in the working age population (aged 15-49) such that those with critical social and economic roles are disproportionately affected.

As to Kloos and Haile Mariam (2000), though the economic and social impacts of AIDS in Ethiopia have not been comprehensively studied and quantified, they are significant and growing. HIV/AIDS results in decreasing of economic growth and preventing the participation of the population in the economy. A World Bank (2000a) as it is quoted in Kloos and Hail Mariam (2000) study estimated that AIDS is already causing a one percent annual reduction in economic growth in Ethiopia, which, together with declining life expectancy and labor force reduction, is systematically undermining the country's efforts to reduce poverty through improvements in health, education, agricultural production, and household food security.

Since HIV/AIDS strikes working-age members of society, it hinders the successful economic development of countries and the profitability of companies. Furthermore, a delayed economic and social impact arises as a result of children losing their parents, and being deprived of basic parenting as well as being forced to leave school prematurely without basic skills.

2.5.4 HIV /AIDS Risk factors

Anyone of any age, race, sex or sexual orientation can be infected with HIV, but there are some special cases where the risk is high. According to MayoClinic.com (2008), the following are identified as high risk to HIV.

- Have unprotected sex with multiple partners,
- Have unprotected sex with someone who is HIV-positive.
- Have another sexually transmitted disease, such as syphilis, herpes, gonorrhea
- Share needles during intravenous drug use.

An estimated 87% of HIV transmission is through heterosexual contact, 10% by mother-to-child transmission and a smaller proportion is thought to be due to traditional harmful practices (HAPCO, 2004), similar to the pattern reported from other African countries. Although empirical evidence for the relative contribution of the major transmission routes has not been obtained anywhere, heterosexual transmission and harmful practices are the major risk factors in Africa.

2.5.5 HIV/AIDS Prevention and Support Mechanisms

Various prevention mechanisms are forwarded by stakeholders. According to WHO(2008), VCT for HIV ,the promotion on Condom use, handling the case of Mother to child transmission (MTCT), and the introduction of Antiretroviral drugs (ART) which significantly delay the progression of HIV to AIDS and allow people living with HIV to live relatively normal healthy lives are some of the prevention and support mechanism .

Provision of Voluntary HIV Counseling & Testing (VCT) is an important part of any national prevention program. It is widely recognized that individuals living with HIV who are aware of their status are less likely to transmit HIV infection to others, and that through testing they can be directed to care and support that can help them to stay

healthy. VCT also provides benefit for those who test negative, in that their behavior may change as a result of the test.

2.5.6. VCT

VCT is a gateway for early access to prevention, care and support services because, it is believed that counseling helps to reduce risk and HIV positive clients can be referred for needed services .Moreover, people can learn whether they are infected, and are helped to understand the implications of their HIV status and make informed decision for the future (Centers for Disease Control and Prevention, 2003), (Ministry of Health, 2007).

As the possibility of getting treatment in HIV begins and when researches on prevention of HIV transmission from mother-to-child become effective in 1990s, the benefits of knowing one's serostatus was mandatory. According to the Family Health International (2005), researches on HIV and prevention showed that VCT could be a cost-effective intervention in developing countries, including in low prevalence settings. Due to this VCT becomes a basic prevention strategy.

The CDC guidelines which were published in 1987 were revised in 1993 to focus on a model of interactive personalized risk reduction, and again in 1994 to emphasize linking standard VCT procedures with treatment goals.

2.5.6.1 The Counseling process

In recent years, voluntary HIV testing, in combination with pre- and post-test counseling, has become increasingly important in national and international prevention and care efforts. Knowledge of serostatus through VCT can be a motivating force for HIV-positive and -negative people alike to adopt safer sexual behavior, which enables seropositive people to prevent their sexual partners from getting infected and those who test seronegative to remain negative.

The UNAIDS/WHO (2007) policy on HIV testing and counseling defines two main categories .They are client-initiated HIV testing and counseling and provider-initiated HIV testing and counseling.

Client-initiated testing and counseling (CITC) also called voluntary counseling and testing (VCT), occurs when people come to a service to find out their HIV status. CITC emphasizes individual risk assessment and, also, counseling that addresses the implications of taking an HIV test and the strategies for reducing risk. Counseling covers prevention both prior to and after receiving test results and, if results are positive, referral to care, treatment and support services.

Provider-initiated testing and counseling (PITC) occurs when HIV testing and counseling is recommended by health care providers as a standard part of medical care to individuals attending health care facilities. The purpose of PITC is to enable specific clinical

decisions to be made and/or specific medical services to be offered that would not be possible without knowledge of the person's HIV status.

The counseling process has three sessions. Pre-test session, Rapid test session and Post-test session

Pre-test information can be provided in the form of individual counseling sessions or in group health information talks and should provide information on: the clinical and prevention benefits of testing; the potential risks, including stigma and discrimination, abandonment or violence; the measures that will be taken to guarantee confidentiality of test results; services that are available in the case of either an HIV-negative or an HIV-positive test result; and the fact that individuals have the right to decline the test.

Post-test counseling for HIV-negative persons should provide basic information that includes an explanation of the test result, of the window period for the appearance of HIV-antibodies and a recommendation to re-test, if appropriate. It should also include advice on methods to prevent sexual transmission and provision of male or female condoms and their use. In the case of injecting drug users, it might also include provision or advice on where to obtain substitution therapy and safe injection equipment and how to use it.

Post-test counseling for HIV-positive persons should provide psychosocial support to cope with the emotional impact of the test result, referral to treatment and care services, disclosure to sexual and injecting partners, basic advice on methods to prevent HIV

transmission, provision of male and female condoms and guidance on their use and other measures.

2.5.6.2 VCT in Ethiopia

Voluntary counseling and testing for HIV has been important in the prevention and treatment of HIV in both developed and developing countries. The Ethiopian government**,** like many other countries in the world, developed an HIV/AIDS policy and VCT guidelines based on World Health Organization (WHO) guidelines and local needs and capacity.

HIV counseling began in Ethiopia in the late 1980s with services expanding throughout the 1990s. In the early 1990s, several national level training programs were conducted for nurses and social workers drawn from all regional hospitals and those in Addis Ababa.

Though there is a consensus on wider access to VCT may lead to greater openness about HIV/AIDS and less stigma and discrimination, most people remain unaware of their HIV status due various reasons. According to MOH (2007) with the development of affordable and effective medical care for people living with HIV, demand for testing is increasing rapidly, creating urgent need to increase access.

The number of VCT service providers, including government institutions, NGOs, and private institutions, increased over the last few years. Governments and international donors are strengthening their technical and financial support to improve quality and coverage of VCT services (UNAIDS 2002).

As it is true in the world, Ethiopia is taking measures to prevent the spread of the virus using prevention mechanisms recommended by WHO /UNAIDS. One of such prevention mechanism is VCT. According to the AIDS Resource Center (ARC), there are 591 VCT centers in the country. Of which 171 (28.9%) are found in Addis Ababa.

2.5.6.3 VCT data and Legality issue

According to the Federal Ministry of Health Guidelines for HIV Counseling and Testing in Ethiopia (2007) research can be made on counseling data and process for the purpose of program development of the prevention mechanism and to improve the quality of services. For this purpose the Minster has formulated a policy.

The policy statements are "All HIV counseling and Testing research shall conform to the relevant legislation and ethical standards of practice set by appropriate research ethical committees at various levels and the Government of Ethiopia promotes and encourages research to improve access and quality of service delivery"(MOH, 2007).

With the above guiding policy frame work researches on VCT data can be conducted to support the counseling process and helps the decision makers to make an informed decision.

With respect to the legal issue of VCT The WHO and MOH VCT guidelines confirmed that, VCT is legal as long as the following three points are fulfilled .Firstly, it should be voluntarily, secondly there should be a pre and post-test counseling, and finally it should

be confidential (MOH, 2007). Because of the implications of being HIV-positive it is important that individuals are counseled if they are going to undergo testing. This should take the form of pre-test counseling, where it is made clear what the implications of testing positive or negative are, and post-test counseling.

VCT client record is made by following the policy requirements .Records are made by using client code, so that ethical issues will be handled and the data will be available for research.

As it is indicated by Abreham (2005), the availability of VCT data in electronic format from CDC support centers, will help to conduct this research. In addition, VCT is not limited to any kind of population group and the data from such service will confidentially represent the mass. This is why this investigation uses the data collected from CDC support VCT centers.

2.6 Review of Related Research

2.6.1 General Applications of Data mining

Nowadays, data are being collected and accumulated at an impressive rate across a wide variety of fields. In response to this huge data, the need for a new tools and techniques to assist humans in extracting novel and useful information and/or knowledge from the large volumes of digital data is mandatory .Such tools that support in extracting patterns from large dataset are the subject of the emerging field of KDD (Fayyad,et al,1996).

Data mining, which refers to a particular step in the KDD process, is increasingly popular because of the substantial contribution it can make to industries. Nowadays many organizations are using data mining to help and manage their tasks (Tibebe, 2005).

The Telecommunications and credit card companies are one of the intensive users of data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area in which they are using data mining to predict the effectiveness of surgical procedures, medical tests or medications and identify patterns from patient history (Two crows, 1999).

2.6.2 Application of Data mining in Health organizations

As it is clearly perceived from the previous discussions, Data mining technology provides a user oriented approach to extract novel and hidden patterns in the data. In health industry, with the widespread use of medical information systems including databases, there is an explosive growth in their sizes. The traditional manual data analysis has become insufficient, and methods for efficient computer assisted analysis are absolutely necessary. The way to extract the knowledge in a comprehensible form from the huge amount of data is the primary concern (Kaur and Wasan, 2006).

According to Kaur and Wasan (2006), it is possible to implement KDD in a health care organization with the help of experts who has good understanding of health care industry.

Health care analyst and policy makers can learn lessons from the use of KDD in other industries and apply KDD to problems of health care industry (Hospitals, Insurance companies, Physicians and Pharmaceutical Companies etc.).

For DeGruy (2000), many health service providers are migrating toward the use of computer based patient records and store a large quantity of patient data on test results, medications, prior diagnoses, and other medical history. This is a valuable source of information that could be better used by employing KDD techniques. KDD can be used for identifying patients who should receive flu shots, identifying patients who should enroll in a disease management program and identifying patients who are not in compliance with a treatment plan.

In relation to the application of data mining in health sector Two Crows (1999), suggested that health care organization must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care.

There are many evidences which approved the application of data mining in different health related problems. Some of these applications are shortly and briefly reviewed as follows.

Data mining is often used in detecting health care fraud. IBM Fraud and Abuse Management System is used for detecting health care fraud and abuse, which ranks as one of the nation's leading law enforcement frustrations (IBM, 2003b).

According to Han and Kamber (2001), the past decade has indicated an explosive growth in biomedical research such as DNA analysis. The causes for some disease and their prevention mechanism were not known. Data mining finding on DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as the discovery of new medicines and approaches for diseases diagnosis, prevention, and treatment.

As to DeGruy (2000), one health maintenance organization used KDD techniques and historical data from other patients to determine which of its enrollees may be at risk for certain diseases. Targeted intervention for these enrollees and diseases makes sense; it keeps the enrollee healthier and lowers the provider's cost of treatment.

The work of Shegaw (2002), with the objective of developing a model that could support predicting child mortality has showed the potential applicability of data mining in health sector.

2.6.3 Applicability of Data mining to HIV and VCT

According to CDC (2000), Data mining can be used to identify clusters of HIV Infected Adolescents and Young Adults Cluster. The work of Tettey, et al (2007), showed that Neural Networks can be used, in an inverse configuration, for the adaptive control of HIV status of individuals to understand how demographic properties (for example educational

level) affect the risk of being HIV positive. Data mining can also be applied in mining complex genotypic features for predicting HIV-1 drug resistance **(**Saigo, et al, 2007).

The work of Abreham (2005), which used VCT dataset, showed the possible applicability of data mining techniques in identifying determinant risk factors for HIV.

As it is indicated in the first chapter of this research work, the work of Abreham (2005) is the first attempt to check the applicability of data mining on HIV/AIDS data collected at VCT centers for the purpose of identifying determinant risk factors of HIV and to find their association rule.

The data set for the investigation were client data set collected from VCT Centers,82 attributes (Columns) and 18646 records (row) were the original size. The data mining task in this work were association rule mining He uses WEKA data mining tools to get APRIORI implementation. From this research the rules were generated and were tested for 16 cycles.

Based on the APRIORI algorithm the best rule that was found in the research are summarized as follows.

- If some one is unmarried and he/she expects negative result, there is a 94% confidence, that the person's result is HIV negative.
- If a person is not married, never use condom, and his/her HIV test is negative there is a 94% confidences with 27.8% support that the client will not have sexually transmitted infections.

- For the HIV positive class, records whose STIHIST is no has also shown positive results with 75.53% support. This means there were 100% confidences that 679 records of the 899 HIV positive totals are also NO STIHIST.

Researches on VCT client data set has been conducted using the traditional statistical methods. In this respect the work of Antenane, Mebiratu, and Solomon (2005) is the one that should be mentioned. They were analyzing Socio-demographic profile and prevalence of HIV infection among VCT clients in Addis Ababa. The data for this study were taken from VCT clients' records in the Kassanchis health center. Data were analyzed using EPI-Info and SPSS.

From their findings, 54.1% of the clients were females and the clients' average age was 27.4 years. Seven in ten of the clients were never married and nearly half of the clients (48.8%) have attended secondary school. Unemployed clients accounted for about two-thirds (66.4%) of the total clients. The majority of clients (88.2%) have had sexual practices in the past. Condom use among sexually active clients was found to be low. Most of the variables included in their regression model had significant effects; however, as indicated, sex, employment status and a history of STIs had considerable and statistically significant influence on HIV infection than any other single factor.

As it is indicated in the first chapter of this research work, one of the major points in the research work is to fill gaps .In addition findings using one method can be verified using other methods and techniques. This will help in solving the problems in any area.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Overview

Before one attempts to extract useful knowledge from data, it is important to understand the overall approach. As suggested by Cios et al. (2007), simply knowing many algorithms used for data analysis is not sufficient for a successful data mining research. The challenge for modern data miners is to come up with widely accepted standards that will stimulate major industry growth. To this end, having predefined methodology in data mining researches is mandatory. This chapter indicates type of data mining research models, methods, tools and procedure followed by the researcher.

3.2 The KDD process model

The knowledge discovery process requires a model that consists of a set of processing steps to be followed by researchers when executing a knowledge discovery task. The model will help to describe procedures that are performed in each of its steps. According to literature in the filed of data mining the most popularized Knowledge discovery process models are the CRISP-DM and Fayyad et al (Cios et al., 2007).

In comparing the models the same author argued that the main differences between the CRISP-DM and Fayyad et al. model lie in the number and scope of their specific steps. A

common feature of all models is the definition of inputs and outputs. In Fayyad et al. the numbers of steps to be followed are nine. It provides detailed technical description with respect to data analysis, but lacks business aspects .In CRISP-DM the number of steps are 6, and has good documentation, divides all steps in to sub steps. This will help to easily identify all necessary details in the knowledge discovery process.

The CRISP-DM model is selected as a knowledge discovery process model for this research because it is non-proprietary and focuses on business issues. Moreover CRISP-DM unlike Fayyad et al. does not describe a particular data mining technique; rather it focuses on the process of a data mining project's life cycle. In addition CRISP-DM is Tool-independent / industry-independent and it is a hierarchical process model

Based on the CRISP-DM model how business understanding, data understanding, data preparation and preprocessing, modeling, analysis and evaluation process conducted are explained as follows.

3.2.1 Methods for Business Understanding

The effectiveness and efficiency of research work is determined by good understanding of domain/business area. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into data mining problem definition, and designs a preliminary project plan to achieve the objectives.

There are a number of VCT centers that provide counseling and testing services. The OSSA VCT center, which is supported by CDC, is the organization selected for this research.

To understand the domain and procedures of VCT for HIV, the researcher did a survey for one month at VCT centers and offices where domain experts are found. In this survey the researcher conduct discussion with three domain experts two from VCT data management section W/o Yewubnes Hailu, Ato Tamirat Ayana , and Ato Haileluel Abegaz from the counseling section .The discussion includes about the nature of the client records about the incorrect values, null values, clusters identified and the model selected. This has helped to understand the ideas, suggestions, and alternatives that the users have on the counseling process. All this helps the researcher to have good relationship with them and tried to communicate the benefit of the research outcome and appreciate their contribution to the research work.

The serostatus of clients can be determined based on their HIV blood test result. The HIV test result can be either positive or negative. Based on the current working procedures test result is used as a class label in the counseling process. Other variables like the age, sex, martial status, history of STI, and others are assessed to check weather they can yield better result in identifying risk factors and determining the serstatus of clients using data mining.

3.2.2 Methods for Data Understanding /Collection

This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets.

To understand the data the researcher produce a data collection and description report as it is mentioned in Two Crows (2005).The following figure shows the original dataset in Excel format.

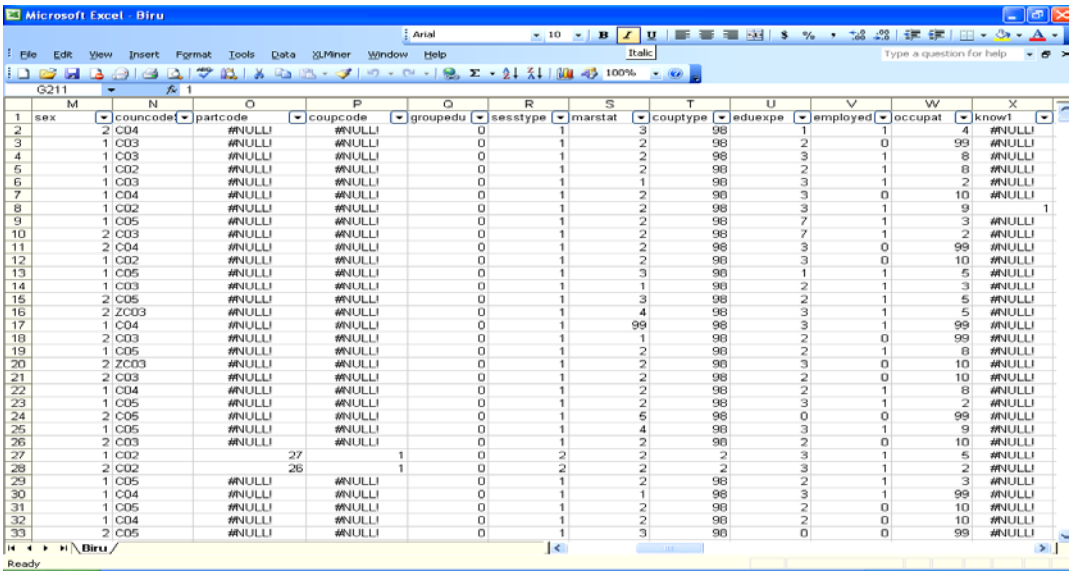


Figure 3. 1: Original dataset in Excel Format

The dataset for this research were in Epi Info software package designed for the global community of public health practitioners and researchers in CDC. It stores VCT client data. A total of 56,468 records (36.8MB) with 82 attributes, in REC file format were

exported to Excel format from the data base and where placed in a CD-ROM. The record was collected for about 7 year, from year 13/03/2002 to 11/06/2008.

As it is seen in Figure 3.1, the dataset where organized in column and rows, where the columns represent an attribute and the row represents single records of clients.

3.2.3 Methods for Data Preparation and Preprocessing

According to Kantardzic (2003), all raw data sets initially prepared for data mining are often large; and may be subject to human error. It is expected to have missing values, distortions, misreporting, inadequate sampling, and so on in the initial data sets. To this end preparation and preprocessing of dataset are critical for data mining researches.

As it is stated by Dasu and Johnson (2003), data preparation is typically the least formalized, the most domain-dependent, and the most time consuming part of the knowledge discovery process. The researcher has taken considerable time in data preparation and preprocessing. In this step the researcher covers all activities needed to construct the final dataset, which constitutes the data that will be fed into Data mining tool in the next step. The following sub steps were done in this research.

3.2.3.1 Data Cleaning Methods

Even if there are huge amounts of data, having completely and correctly filled data may be relatively small. As it is argued by Kantardzic (2003), some of the data-mining methods accept missing values and satisfactorily process data to reach a final conclusion,

other methods require that all values be available. In this research the cleaning process is conducted manually and using tool (WEKA).The tool was used in handling discritization and normalization process. Incorrect values were handled manually.

Cleaning the data manually was found important and effective preprocessing mechanism. This is due to the nature of the dataset and the database. The Epi-Info database accepts incorrect values. If we let this problem to be processed by tools, it is not recognized as a problem.

As it is indicated in Figure 3.1 original dataset has a number of missing values for all attributes. There are different alternative in handling missing values and incompletes. For Han and Kamber (2001), one and most commonly used methods are filling the attributes quickly without too much computation by replacing all the missing values with the arithmetic mean or the mode with respect to that attribute. By adopting the methods above and incorporating the ideas of experts the researcher has performed the data cleaning task as follows.

- Missing value for continuous/numeric value like age are replaced by the mean value of the field. For nominal variable ,the modal (most frequent) value will be filled
- In some cases, the researcher with the domain expert manually examine samples that have no values and enter a reasonable, probable, or expected value, based on a domain experience.

- Outliers and noisy field values were handled and different action was taken. For records which have observed having very great difference from the range of values for the specified attributes, a correction will be made. This might happen due to the error made when entering the data in to the database by the users. At this time correction was made by replacing the values with logical estimates suggested by experts.
- Records or variables with high noisy values were removed.

3.2.3.2 Data Integration and Transformation Methods

The original VCT client data set values of some attributes selected for the research purpose were represented using numerical codes. Processing this numeric code values may result in problems during interpretation. To avoid such problems the researcher has done transformation of attribute values when necessary.

For the transformation process the researcher has used the result of previous research by Abreham (2005) and, the Ethiopian HIV Counseling and Testing Record form of OSSA (See Annex A).

Many attribute/fields in the dataset take more than one value. For example the attributes that describe occupation title of the client (OCCUPAT) and most important reason here today (REASHERE) have 13 and 22 unique values respectively. This will have a great impact on the performance of the algorithm selected.

As it suggested by Cios et l.(2007) discretization and concept hierarchy are a necessary preprocessing step, not just a tool for reducing the data. For the decision tree algorithm selected in this research, reducing the number of values of attributes increases the performance. Because the same author argued that, if the number of values of the attributes/features is huge, model building for such data can be difficult and/or highly inefficient. Discritization will help to minimize such problems

A concept hierarchy is defined for the categorical data. Categorical data are discrete data. Categorical attributes have a finite (but possibly large) number of distinct values, with no ordering among the values. For example in VCT dataset, educational back ground, Occupation title, and martial status are categorical attributes with no ordering among the values. The formation of new concepts is essential.

For Han and Kamber (2001), there are several methods for the generation of concept hierarchies for categorical data. One of such method is defining the concept hierarchy by user or expert by specifying a partial or total ordering of the attributes at the schema level.

To normalize attributes value the Min-max normalization as suggested by Han and Kamber (2001) is used. Min-max normalization performs a linear transformation on the original data and it used because it preserves the relationships among the original data values.

3.2.3.3 Data Reduction and feature selection methods

Irrelevant attributes has a damaging effect on machine learning schemes. Learning with attribute selection, helps to eliminates irrelevant and consider only most relevant attributes. Some Learning methods themselves try to select attributes (decision tree) appropriately and ignore irrelevant or redundant ones, but in practice their performance can frequently be improved by pre-selection (Witten and Frank, 2005). Based on this the selection of relevant attribute is done working with expert

As it is argued by Witten and Frank (2005) in many practical situations there are far too many attributes for learning. There should be a mechanism of selecting relevant attributes. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean.

Selection of relevant attributes is important and crucial in data mining. In this research work the reduction and selection of attribute will be done first manually in consultation with experts by considering the significance and importance attributes to the objectives of the research work.

Once the relevant attributes are selected by experts, further selection and comparison will be made by using Information gain on WEKA data mining tools. Dimension reduction helps to balance the training dataset so that the algorithm will have a chance to learn from every group of records.

3.2.4 Methods for Modeling

One of the basic activities in Data mining is model building. As it is mentioned in chapter two of this research work at very general level, models of data mining is either prediction or description.

The description task, unsupervised learning, is used as a preprocessing stage to identify pattern classes for subsequent supervised classification. For this purpose a number of clustering algorithms are available. A single algorithm or approach is not adequate to solve every clustering problem (Jain et al., 1999).

For Han and Kamber (2001), the choice of clustering algorithm depends both on the type of data available and on the particular purpose of the application. Trying several types of algorithm on the same data to see what the data may disclose is essential. Based on this from the variety of available clustering algorithms the K-Means and expectation maximization (EM) algorithms are selected by the researcher.

The *k*-means algorithm is selected for the clustering task because it is most successfully used on large data sets and it is simple to implement and computationally attractive. In addition this algorithm time complexity is linear (jain et.al, 1999).

Even if the K-means has advantage, it has a problem related to the selection of the initial cluster number (K). Several strategies can be used to overcome this disadvantage. Trying the cluster assignments of using many trials by varying random seeds is one method. Another alternative is to use an algorithm such as the EM algorithm that is not subject to

this disadvantage. The EM algorithm is selected by the researcher to compare results of K-means and by pass the problem associated with it.

The EM algorithm is selected from the other because it offers multiple advantages in comparison to k-means clustering:

- Requires one database scan, at most.
- Outperforms sampling approaches
- It implements a statistical model that, when appropriate, shows optimal results.
- The implementation of EM assumes all attributes to be independent and normally distributed.

In this research work the classification model was built by using the decision tree algorithm. This is because decision tree is used to predict categorical variables Two Crows Corporation (2005). More over decision tree algorithms are used because

- They are easy to understand
- They are easily converted to a set of production rules
- They can classify both categorical and numerical data, but the output attribute must be categorical
- There are no a priori assumptions about the nature of the data.

Kantardzic (2003) argued that decision trees make few passes through to the data. In most cases there is only one pass for each level. In addition decision tree are effective for classification when there is a class label or predictor variables. From the available

decision tree algorithms that are supported by the WEKA tool J48 and Random Tree are analyzed and compared.

Artificial Neural network is another algorithm used for classification purpose. The main advantages of a neural network are its robustness to data noise, its capacity of running in parallel as well as its capacity of approximation any function. The more serious disadvantage of a neural network is that the model (e.g.: the weights of the neuron connections) is incomprehensible for a human and no business knowledge can be extracted from it.

3.2.4.1 Adjusting parameters for modeling

Larose (2006), for learning models, if one of the target variable classes has much lower relative frequency than the other classes, balancing is recommended. The same author explained this problem by considering a fraud classification model which has 100,000 transactions, only 1000 of which are fraudulent. By considering this example the classification model could simply predict no fraudulent for all transaction and achieve 99% accuracy. However this model is useless.

The above mentioned problem for Pyle (2003) is called the naïve prediction rate (or naïve error rate). The amount of other data simply swamps the relatively low level of information that is present. In order to get the needed information exposed to the tool, the data set has to be adjusted.

The researcher identifies similar problem like fraud example mentioned above in VCT client dataset in which 87% are HIV negative and 13% HIV positive. In this situation, the researcher balance the training dataset by using purposeful stratified random sampling technique and considers a total of 14793(7111 HIV positive and 7682 HIV negative).

The selection of proportional sample is made from the HIV negative client record by considering the gender of clients. For this purpose to consider proportional number of male and female clients the researcher filter records by sex and calculate k, so that every kth row beginning from the first row will be selected. The selections of samples begin from initial row. To have 54% male and 46 % female sample the researcher uses the Excel function and retrieved 4408 Male and 3274 Female (Total of 7682) clients.

3.2.4.2 Training Methods

The performance of different algorithm can be measured after a serious of training. Data mining tools WEKA has the following training set, supplied test set, cross-validation and percentage split training options. From these options the researcher selects training set test options for the clusterer evaluation and cross-validation test option for classification task.

According to Olson and Delen (2008), 10-fold cross validation is good training option because it does not require more data compared to the traditional single percentage split (2/3 training, 1/3 testing) experimentation.

The main advantage of 10-fold (or any number of folds) cross validation as it is explained by the same author is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as holdout sample.

The problem associated with this method is that, one needs to do the training and testing for *k* times (*k* = 10 in this study) as opposed to only once. Even though this methodology is rather time consuming, for some small datasets it is a viable option.

For the clustering phase selecting the cluster mode is essential. In WEKA to evaluate the clustering there are four different modes. They are Use training set (default), Supplied test set, Percentage split and, Classes to clusters evaluation.

According to Witten and Frank (2005) for probabilistic clustering methods (like EM), percentage split and supplied test set evaluation mechanism can be used. In addition in this method WEKA measures the log-likelihood of the clusters on the training data. The larger this quantity, the better the model fits the data. Increasing the number of clusters normally increases the likelihood, but may overfit. The researcher used the percentage split cluster mode and the percentage were arranged manually to use 80% of the record for the training and 20 %for the testing.

3.3 Methods for Analysis and Evaluation

In this step, the researcher analyzed the result of the algorithms (classification and clustering) by a selected tool. The results of the algorithms are also compared with the ideas of experts.

The analysis and evaluation was conducted by observing the confusion matrix. The confusion matrix is simply a square matrix that shows the various classifications and misclassifications of the model in a compact area. The columns of the matrix correspond to the number of instances classified as a particular value and the rows correspond to the number of instances with that actual classification.

The researcher used exploratory data analysis technique to analyze the results. This technique is used to present results in graphs and tabular formats. For model evaluation and selection even if the above parameters can be used, the suggestion, choice and opinion of domain experts and users are taken in to consideration.

As it is clearly stated and suggested by Larose (2005) confluence of results are necessary in model selection. The main idea in model selection is that the researcher and analyst should not depend on one kind of methods. Based on this the researcher selects models by comparing different algorithms, users comment and suggestions.

3.4 Data Mining Tool Selection

As it is suggested by Nisbet (2004), the choice of a data mining tool is not an easy task. The same author argued that, the best tool suite for some one may not be the most advanced tool, or the one that gives the greatest accuracy in prediction. More important than all of these things is identifying the tool suite that is easy to use, provides acceptable accuracy (even though not the highest accuracy available), and able to perform all the common tasks in a data mining project

The most widely known commercial data mining tools are Clementine from SPSS, Enterprise Miner from SAS, Intelligent Miner from IBM and Statistica from StatSoft. On the other hand there are open source data mining tools such as WEKA, TANAGRA, and Rapid Miner

One of the major problems faced by the researcher at this stage was getting commercial available tools. This is because of some of these tools are available free for downloading and evaluation without charge only for some period of time. Some are really expensive. Having much effort and sharing previous experiences of researchers on data mining, the researcher was able to get WEKA, TANAGARA, Rapid miner data mining tools.

3.4.1 WEKA

According to Whitten and Frank (2005) WEKA, which stands for Waikato Environment for Knowledge Analysis, was developed at the University of Waikato in New Zealand.

The system is written in Java and distributed under the terms of the GNU General Public License.

WEKA can handle all the standard data mining problems like regression, classification, clustering, association rule mining, and attribute selection. Moreover it is an open source data mining tool. This tool is available for free and can easily be downloaded from site http://www.cs.waikato.ac.nz/~ml/weka/.

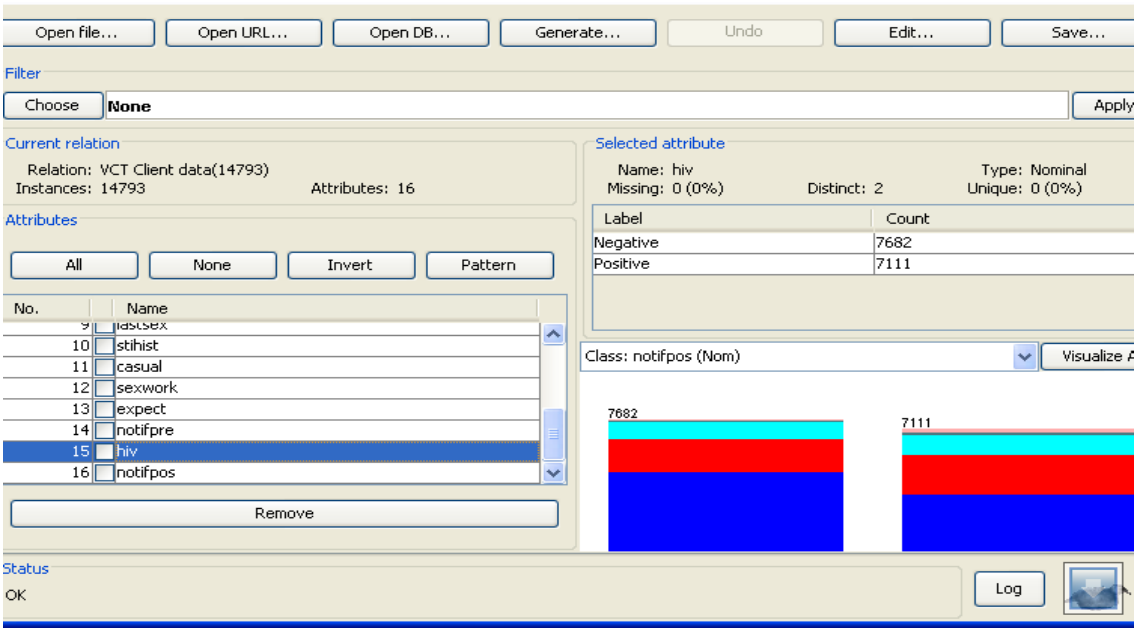


Figure 3. 2: WEKA Interface in VCT dataset

3.4.2 TANAGARA

TANAGRA, which was developed in France, is a free data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area.

TANAGRA provides most of all the categories of data mining methods, like regression, classification and clustering. The major problem in this tool is it is highly sensitive to variable types.

In relation to its advantage TANAGRA can handle spreadsheet data set and WEKA file formats,(.arff).Due to this it is possible to compare results of the two application. The main functionality of TANAGRA is data visualization, descriptive statistics, Clustering and supervised learning.

3.4.3 Rapid Miner

Rapid Miner, Previously known as YALE, is a data mining suite which makes a wide range of techniques available. It builds on the WEKA data mining tool and adds a number of useful (and pretty) visualization methods. It is available free under a GNU General Public License (GPL) or under a paid for proprietary license which can allow commercial redistribution. Even if rapid miner has much application for data mining, some of the functionalities are not found in its free version. Moreover the researcher was not able to practice and explore possible functionalities.

In this research work the researcher tried to get and have different data mining tools. From the freely available data mining tools the current version of WEKA 3.5.8 was selected as tool because it can support classification and clustering task, has number of algorithms, and the tool support and can load input file format that are in Excel. In addition the tool is easy to use.

CHAPTER FOUR
EXPERIMENTATION AND DISCUSSION OF RESULTS

4.1 Overview

There are a number of data mining algorithms that are found useful for automatic classification and clustering of data. Some algorithms might work better than others while running one type of data as compared to the other. Thus finding the best type of algorithm is an interesting and time consuming work.

Two major data mining tasks,(Clustering and Classification) are going to be explored. For the clustering task the K-Means and EM algorithms and for the classification task decision tree and ANN are going to be analyzed. Applying the CRISP-DM process model,the exprimentaiuon begins by the clustering phase.The clustering phase is followed by the classification phase.

Both the classification and clustering experiment are conducted by using different combination of attributes and parameters. Moreover the classification task is performed in two sub phases using two different kinds of class labels(One class label selected manually and the other is the class label identified in clustering phase).

As it is indicated in the previous chapter the tool that is used for the classification and clustering phases is WEKA. The classification models are built using decision tree and ANN algorithms of WEKA, and clustering models are built using the K-means and EM algorithms.

4.2 The data selection process

The data set for this research were obtained from OSSA VCT center which is supported by CDC. The data was stored in Epi Info data base.

The data collection process begins by exporting the data from Epi Info data base. In doing this the records were found in Excel format. Since the database is used entirely to handle VCT client data, all stored VCT client record values from 13/03/2002 to 11/06/2008 where exported from the data base.

4.2.1 Basic data description

The dataset were organized in column and rows, where the columns represent an attribute and the row represents single records of clients of VCT. The original dataset has 82 attributes (Columns) and 56468 records (rows). Attribute names, Data type, number of unique values they take, and number of missing value of 14 selected attribute is indicated in Table 4.1.

No	Attribute	Data Type	Values	Description	Number of Missing values
1	AGE	Numeric	Continuous numerical values	The age of the client	25
2	CASUAL	Numeric	Continuous numerical values	The number of casual partners of the client	26
3	EDUEXP	Numeric	9 Unique categorical values	Educational level of the client	47
4	EMPLOYED	Numeric	2 Unique categorical values	Employment condition of the client	59
5	EVERHAD	Numeric	2 Unique categorical values	Previous sex experience of the client (with penetration)	38
6	EXPECT	Numeric	4 Unique categorical values	A test result that a client expects	38
7	HIV	Numeric	3 Unique categorical values	The HIV test result of the client	72
8	LASTSEX	Numeric	4 Unique categorical values	The condom use of the client during the last time had sex	44
9	MARSTAT	Numeric	7 Unique categorical values	Martial status of the client	32
10	OCCUPAT	Numeric	13 Unique categorical values	Occupation title of the client	226
11	REASHERE	Numeric	5 Unique categorical values	Most important reason the client is here today	58
12	SEX	Numeric	2 Unique categorical values	The gender of the client	8
13	SEXWORK	Numeric	5Unique categorical values	Is the client commercial sex worker	31
14	STIHIST	Numeric	4 Unique categorical values	STI history of the client	48

Table 4. 1: Data description of selected attributes

4.2.2. Data Preprocessing

As it is seen in the Table 4.1, much of the raw data contained in VCT dataset are incomplete and noisy. There are fields that are obsolete or redundant and containing missing values. Such unimportant (That are not needed for this research) attributes are removed and the remaining attributes are considered for further preprocessing.

4.2.2.1 Data Cleaning

To clean the VCT client dataset and prepare for the mining purpose, initial feature selection is done as it is indicated in the methodology part of this research work. The selection of relevant attributes is made by working with experts. The expert and the researcher select relevant attributes which are displayed in Table 4.1. Once attribute/fields are selected the cleaning process is made to handle problems related with missing and incorrect /noise values.

Missing value was observed almost in all attributes. Usually missing values are replaced either by mean value or modal value for numeric and nominal attributes respectively.

- For numerical variables like age, the missing value is replaced by the mean age. This is done by the tool. The WEKA replaces missing values by mean value.
- For all nominal variables like MARSTAT, EDUEXP, EMPLOYED, OCCUPAT, the modal value is filled.

To handle outliers and noisy field values, the researcher has done manually the following task:

For some attribute values data was entered incorrectly in to the database. This will create noise and outlier in the data set. For example the VCT client record begins in 2002. But a record is found for the year 2001.For some other attributes whose value unique categorical value, incorrect values other than the mentioned unique values where obtained. Such problems are handled and correct values are substituted based on the methods described in the previous chapter. The following table displays some of the mentioned cases

Attributes	**Values It takes**	**Incorrect stored Value**	**Corrected values**
MARSTAT	0-6,99	98	99
EDUEXP	0-7,99	36,98	99,Modal
EVERHAD	0,1	3,4	Modal value
STIHIST	0,1,97,98	3,9	Modal value

Table 4. 2:Sample of Incorrect Values

From this case it can be concluded that the Epi Info database allows errors that are made by data encoders (users). Due to these records contained incorrect values. In this case the researcher does the following in consultation with experts.

- When the attributes value is from 1 to 6 and if the value in the database is other than this, the wrong value will be substituted by the modal value.
- When the exact value of the attribute is 97 or 98 or 99 and the incorrect value is like 79,89 the researcher corrects by estimating the nearest possible value ,79 by 97, 89 by 98 and so on.

- For age attribute some of the values were 0, in this case the researcher decided to substitute with the mean age.
- For year attribute with incorrect value of 2001, it is replaced by modal value of year.

4.2.2.2 Data Encoding and Decoding

For the decoding and encoding process the researcher has used the work of Abreham (2005) which used similar dataset and the Ethiopian HIV Counseling and Testing Record form of OSSA (See ANNEX A).The form is filled by the counselor at the counseling time. Data encoders enter the data based on the description that exists on the form. The following tables show some of the transformed attribute values.

Attribute Name	**Original code**	**New Code**
SEX	1 and 2	Male for 1
		Female for 2
EXPECT	0,1,97,98	Negative for 0, Positive for 1, Don't know for 97 , and Not applicable for 98
HIV	0,1,98	HIV- for 0,HIV + for 1,and Not applicable for 98

Table 4. 3:Sample of Attributes values decoded

In this research discritization and concept hierarchy definition activities performed are based on the ideas obtained from domain experts and methods mentioned in the previous chapter. Attributes values which were discritized and summarized by higher concepts are presented as follows.

Two continuous attributes AGE and CASUAL are discritized by using WEKA tool. Age is discritized by using the bin method in WEKA. The number of Bin was adjusted to be 5. Similarly CASUAL is discritized.

The MARSTAT,EDUEXP,SEXWORK,OCCUPAT,and REASHERE attributes new concept hierarchy is presented as follows.

Marital status (MARSTAT): The MARSTAT attributes which describes the martial status of the client has the following values after decoding the numeric code in to symbols. .Married, Never married, Separated, Divorced, Widowed, Living together and Other.

- Married and Living together values were represented by Married.
- Separated and divorced were represented by one value (Divorced).

Code	Original	New
1	Married	Married
2	Never Married	Single
3	Separated	Divorced
4	Divorced	Divorced
5	Widowed	Widowed
6	Living Together	Married
99	Other	Other

Table 4. 4: MARSTAT Attribute

Educational background (EDUEXP): The educational background of clients was recorded by 8 distinct values.

- Illiterate, able to read and other were represented by one value (Other).

- 7-12 in new curriculum, 11-12 academic, and 11-12 vocational were represented by Sec_ New.

Code	Original	New
0	Illiterate	Other
1	Able to read	Other
2	1-6 primary	Primary
3	7-12 old	Sec_Od
4	7-12 new	Sec_New
5	11-12 A Lev	Sec_New
6	11-12 Voc	Sec_New
7	>12	Tertiary
99	Other	Other

Table 4. 5:EDUEXP Attribute

Is client commercial sex worker (SEXWORK): This attribute were represented by five distinct values.

- Yes and counselor thinks yes are represented by Yes
- Didn't know and Not applicable are represented by Not_ appl

Code	Original	New
0	No	No
1	Yes	Yes
2	Counselor thinks yes	Yes
97	Didn't know	Not_appl
98	N/A	Not_appl

Table 4. 6: SEXWORK Attribute

Occupation title (OCCUPAT): This attribute is used to indicate the occupation of the clients who are employed. In the employed attribute if the client answer is yes, his occupation title will be recorded in the occupation title. Clients who are not employed were not expected to respond to the occupation title values. According to the information obtained from domain experts, unemployed clients record in relation to occupation should be represented by the attribute value "Other". But, in the record it is represented by Student and Housewife.

- The researcher and experts concluded to represent Students and Housewife by Unemployed (Un_ Empl).
- The other values are replaced by new concepts as it is shown in the following table.

Code	Original	New
1	Managers, officials	Officials_Proff
2	Professionals	Officials_Proff
3	Technicians,Ass.professionals	Officials_Proff
4	Clerks	Clerk_Sales
5	Sales, service, shop, market	Clerk_Sales
6	Skilled agr. and Fishery	Machine_Craft
7	Crafts and trades	Machine_Craft
8	Machine operators	Machine_Craft
9	Unskilled occupation	Unskilled_Occu
10	Student	Un_Empl
11	Housewife	Un_Empl
12	Armed force	Armed_Force
99	Others	Other

Table 4. 7: OCCUPAT Attribute

Most important reason here today (REASHERE): Most important reason here today attribute is used to describe the reason why the client has visited the VCT client center at OSSA. It is represented by 22 numeric codes each representing unique categorical values like most attributes in the dataset. Using all 22 values for one attribute will have an impact on the algorithms as described in the literature and methodology part of this research. Due to this, representing some of the values by some other higher concepts is done as it is indicated in the following table

Code	Original	New
1	Client risky/Hard risk	Hard _Risk
2	Partner risky/hard risk	Hard _Risk
3	Not trust partner	Other
4	Symptoms	Hard _Risk
5	Premarital	Planning
6	Martial reunion	Planning
7	Family planning	Planning
8	Visa applicants	Planning
9	Referred	Sec_Test
10	2nd test (window)	Sec_Test
11	Confirm positive result	Sec_Test
12	Get results prev.test	Other
13	Need counseling	Other
14	Test before pregnant	Planning
15	Pregnant must know	Planning
16	Plan for future	Planning
17	Death/illness of the partner	Hard _Risk
18	Occupational exposure	Hard _Risk
19	Other blood/fluid exp	Hard _Risk
20	Sexual assault	Hard _Risk
21	Preliminary ART	Other
99	Other	Other

Table 4. 8: REASHERE Attribute

4.2.2.3 Data Reduction and Feature Selection

Insight from a domain expert is essential in data reduction and feature selection process. Because expertise in the VCT for HIV clients is not trivial, their involvement in the preprocessing step is crucial. Candidate variables were selected on the basis of their influence on the counseling process and based on the research objectives. The reduction process applied on attributes and values of the dataset is described as follows.

- Attributes like syphilis, refer 1, refer 2, refer 3… (attributes which indicates where the client is referred to), and Know1, know2… (attributes which indicates how client hear about the services of VCT in that organization) were removed because they contain too many missing values.
- Attributes which indicate Address (Country, Region, woreda, previous locations were dropped).This is because the dataset was taken from one organization that is found in Addis Ababa.
- Attribute related with some security codes given to the clients and counselor are also omitted.
- The kind of service given by the VCT center at OSSA is eliminated based on the suggestion given by experts.
- For efficient model building balancing the dataset is necessary. As it is indicated in methodology part of this research from a total of 56468 records 87 %(49357) are HIV negative and 13% (7111) HIV positive. As it is indicated in methodology part appropriate samples are taken. Based on this a total of 7682 records were selected fro HIV negative clients.

By using the data preparation and preprocessing methods it was possible to reduce the size of the dataset from 56,468 to 14793 records and from 82 attributes to 14 attributes.

4.2.2.4 Machine understandable format

Raw data can be stored in several formats, including text, Excel or other database types of files. Converting the data in to a format understandable by the selected tool should be performed for further preprocessing if needed.

In the WEKA tool for example, the data should be stored in the Attribute-Relation File Format (.arff format) as the data type of the attributes. The system does not automatically classify the attribute as being real or categorical.

In this regard the researcher prepared the dataset in Excel format, and loads the data set to the WEKA tools as follows. The dataset in Excel format was opened and it was saved as comma separated value (CSV) format. WEKA has CSV file loader and can load it to the tool by converting it into .arff format

The major task in this chapter is the experimentation and discussion of results based on the methodology selected for the research in the previous chapter. Applying the CRISP-DM process model, the experimentation begins by clustering sub phase. The clustering sub phase is followed by the Classification analysis.

In the first part of the classification experiment, classification task is performed by using the class label identified in the process of business understanding. In the second part after the identification of automatic cluster labels, new classification experiment is conducted. Final comparison of the two classification experiment is performed.

As it is indicated in the previous chapter the tool that is used for the classification and clustering phases is WEKA. The classification models are built using decision tree and Multilayer perceptron (ANN) learner algorithms of WEKA and clustering models are built using the K-means and EM algorithms.

4.3 The clustering sub phase

4.3.1 Initial Clustering Analysis: Feature Selection

The main goal of the cluster analysis is to define groups of similar VCT clients and to see how these grouping ultimately affect the classification outcome. According to the literature and VCT experiences, having sex with penetration, history of sexually transmitted infections, martial status, occupation title, condom use last three months, used condom last time had sex, number of casual partners, and age of the VCT clients play an important role in determining clients HIV test result.

The clustering experiment was conducting using all attributes indicated in table 4.1.Step by steps attribute which can yield meaningful cluster are identified. After selecting these

clustering features with the help of domain expert and literature the researcher identifies some important issues that should be addressed in performing the cluster analysis.

- How important will the clustering consider the selected attribute in determining the similarity between VCT clients?
- How will different centriod initialization affect the resulting cluster?
- What are the possible strategies that can be employed to address centroid initialization problem?

4.3.2 Clustering Experiment one

For the first experiment a total of 56468 records with 14 attributes were processed by latest version WEKA 3.5.8 data mining tool. As it is indicated in the preprocessing part of this research the WEKA tool holds everything in memory, in order to avoid memory leakage during experimentation, first the JVM heap size was changed from the default 128MB to 512 MB. then the experiment was carried out by machine with 2 GB RAM capacities and core 2 Duo @ 1.86 GHz CPU processor speed.

4.3.2.1 K-Means algorithms

To test weather there is a natural cluster in VCT client dataset, the K-Means algorithm parameters were adjusted using the default values (K= 2, Seed=10) .For this experiment 14 attributes (AGE, SEX, MARSTAT, EDUEXP, EMPLOYED, OCCUPAT, REASHERE, EVERHAD, LASTSEX, SEXWORK,STIHIST, CASUAL, EXPECT, and HIV) are used . The results are summarized in the Table 4.9

Attribute	Clustered instances					
	56468(100%)	11902(21%)	9837(17%)	11046(20%)	13770(24%)	9913(18%)
	Full data	Co	C1	C2	C3	C4
AGE	26.1782	24.6607	28.5488	20.7941	30.233	24.1511
SEX	Male	Female	Male	Female	Male	Female
MARSTAT	Single	Single	Single	Single	Single	Single
EDUEXP	Sec_Old	Sec_Old	Tertiary	Sec_New	Sec_Old	Sec_New
EMPLOYED	Yes	Yes	Yes	No	Yes	No
OCCUPAT	Other	Unskilled_occ	Officials_Proff	Un_Empl	Clerks_Sales	Un_Empl
REASHERE	Hard_Risk	Hard_Risk	Planning	Planning	Hard_Risk	Hard_Risk
EVERHAD	Yes	Yes	Yes	No	Yes	Yes
LASTSEX	No	No	No	Not_appl	No	No
STIHIST	No	DonotKNnow	No	Not_appl	NO	DonotKNnow
CASUAL	0	0	0	0	0	Negative
SEXWORK	Not_appl	Not_appl	Not_appl	Not_appl	Not_appl	0
EXPECT	DonotKnow	DonotKnow	DonotKnow	Negative	DonotKnow	Negative
HIV	Negative	Negative	Negative	Negative	Negative	Negative

Table 4. 9: Partial View of K-means out put

As it is indicated in Table 4.9,For the K-means clustering algorithms (K=5), in order to see what patterns are discovered the researcher used the dataset and the output of the WEKA tool along with the suggestions of domain experts. According to domain experts' observation, they were not able to see meaningful clusters specially those clusters that include both HIV-negative and HIV-positive. Since the majority of the records are HIV-negative, the tools do not have the chance to learn about the entire dataset. Table 4.9 displays the output of the tool and by using the modal value in the case of categorical attributes.

4.3.3 Clustering experiment Two

As it is indicated in the methodology part of this research, the dataset was balanced and the second experiment was conducted using 14793 records with 14 selected attributes.

The dataset was tested using the K-means and EM algorithms by WEKA data mining tool.

In the first case all 14 attributes were used. Secondly using 10 attributes (ignoring 4 attributes which are LASTSEX, STIHIST, CASUAL, SEXWORK), and finally by using 7 attributes ignoring AGE, EMPLOYED, LASTSEX, STIHIST, CASUAL, SEXWORK, EVERHAD).

This grouping is made based on the nature of the dataset in the database. According to domain experts, LASTSEX, STIHIST, CASUAL, SEXWORK attributes are a key attributes in determining the serostatus of clients. But when we see the record all this attributes are highly skewed to one value, usually No in case of LATSEX, STIHIST, and SEXWORK and zero for CASUAL.

When instances are placed in one of the clusters, the number of instances or the percentage they cover should be considered. If empty clusters are generated during the process, or if the membership of one or more of the clusters falls below a given threshold, the clusters with low populations should be reseeded at new points by rerunning the EM algorithm again or that cluster should be rejected. Based on this the researcher set a size metrics for cluster. For a cluster to be considered, it should contain at least 15%of the record instances.

4.3.3.1 The K-means algorithm

The K-means means clustering, as it is stated in the literature part of this research, is one of the well-known methods of assigning cluster membership by minimizing the differences among items in a cluster while maximizing the distance between clusters. In WEKA there are two parameters that should be adjusted manually. These are the number of cluster K and the number of seed.

Attributes	K=2		K=3			K=4			
	Modal Value		Modal Value			Modal Value			
	C0	C1	C0	C1	C2	Co	C1	C2	C3
SEX	Female	Male	Female	Male	Female	Female	Male	Male	Female
MARSTAT	Single	Single	Single	Single	Single	Single	Single	Single	Single
EDUEXP	Sec_Old	Sec_Old	Sec_Old	Sec_Old	Sec_New	Sec_Old	Sec_Old	Tertiary	Sec_New
OCCUPAT	Other	Clerks_sales	Other	Clerks_sales	Un_Empl	Other	Clerks_Sales	Officials_Proff	Un_Empl
REASHERE	Hard_Risk	Hard_Risk	Hard_Risk	Hard_Risk	Planning	Hard_Risk	Hard_Risk	Hard_Risk	Hard_Risk
EXPECT	DonotKnow	DonotKnow	DonotKnow	DonotKnow	DonotKnow	DonotKnow	DonotKnow	Negative	DonotKnow
HIV	Negative	Positive	Positive	Negative	Negative	Positive	Negative	Positive	Negative
No.Instance	8331(56%)	6462(44%)	7668(52%)	4770(32%)	2355(16%)	6742(46%)	4391(30%)	1338(9%)	2322(16%)

Table 4. 10: Partial View of K-Means for K=2,3,and 4

The cluster seed is used to specify the seed number that is used to randomly generate clusters for the initial stage of model building. By changing this number the initial clusters are built and model built by different seeds are compared. The cluster seed number is changed until the clusters that are found do not change that much.

In this K-means cluster analysis the researcher first used the default parameters of the tool for the initial run, and changes the parameters to 3 and 4 for K and 100 and 10000 for the cluster seed. After observing the out put of the clusterer, and analyzing the parameters, the analysis of the out put for the cluster seed 1000, and the three different numbers of clusters are presented in Table 4.10 above.

As it is indicated in Table 4.10, the K-means algorithm generates clusters for different value of K. These clusters are represented by the modal values. The cluster formed by different value of K are not that much different. Cluster formed by K=3, and K=4 are almost similar. The number of instances under cluster index of C2 in K=4 are 9%. This number is not a sufficient to build a model .If this cluster index is dropped, both K=3, and K=4 can be represented by one cluster .Modal values of attributes for this two clusters are also similar. Therefore the researcher drop cluster formed by K=4.

In comparing K=2, and K=3, the majority of HIV positive clients are under C1 in K=2 and this cluster includes more number of male clients than female. Before deciding which cluster is good, the researcher does other experiment using expectation maximization algorithm**.**

As it is indicated in literature part, one of the draw backs of K-means algorithm is the initialization of K. The initial number of cluster selected by the K-Means algorithm can have a significant effect on the resulting clusters. It can be just as important to properly initialize the clusters as choosing the number of clusters. The other problem in relation to K-means is it is not good in handling categorical or nominal attributes.

Several strategies can be used to overcome this disadvantage. Analyzing the cluster assignments of several trials of the algorithm using varying random seeds is one of the methods and often implemented strategy. Another alternative is using an algorithm such as the EM algorithm that is not subject to this disadvantage.

By default, EM selects the number of clusters automatically by maximizing the logarithm of the likelihood of future data, estimated using cross-validation. Beginning with one cluster, it continues to add clusters until the estimated log-likelihood decreases. The cluster numbers generated randomly by EM are considered as far as they satisfy the minimum threshold set. Because some times over fitting may happen. The EM algorithm analysis using WEKA on the same VCT dataset is presented in the following section.

4.3.3.2 EM cluster Algorithms

In WEKA, there are four cluster modes. The researcher chooses the percentage split and arrange the parameter 80% for training and 20% for testing. The other parameters that are expected to be set manually are the number of iteration and seed number.

The EM cluster algorithm experiment were carried by two phase. In the first phase, the algorithm was run by using the default (number of cluster to be determined by the algorithm it self).In the second phase the number of cluster was determined manually by the researcher.

Phase 1: Instruct EM to determine a best number of clusters by setting the number of clusters to be -1. (-1 for don't care) case:

In this phase the number of records and attributes selected and used for the K-means were used. The cluster seed and number of iteration are set to be 100.Other parameters where taking the default value of the tool.

From the output the number of clusters and instances assigned to each clusters are seen in the above window. Based on the out put of this run:

- The number of cluster generated were 18
- Cluster index 4 contains the maximum number of instances 300(10%)
- The minimum number of instances was observed in cluster index of 10, 12 and 17 which is 50(2%).

Cluster Index	12	17	10	7	11	9	13	8	6	14	5	1	15	2	0	3	16	4
Clustered Instances	2%			3%			4%	5%		6%	7%			8%		9%		10%

Table 4. 11: EM cluster out put for unknown K

From the above table (Table 4.11), the numbers of instances assigned in each cluster are very small. It is difficult to have a model for cluster that is represented by clusters that contains small number of instances. Due to this the result of the EM algorithm output is not considered by the researcher for further analysis. The researcher decided to conduct the EM experiment again by setting the number of cluster manually.

Phase II: Keeping other things constant as it was in the first phase, the researcher set the number of cluster size (K) that should be generated. Initially K was 2, then 3, and finally 4.

The initial selection of K to be 2 was based on the business understanding. In the existing system VCT centers have class label that are considered as clusters (HIV test result which is either positive or Negative).

The out put of the EM cluster is displayed in Table 4.12, for K=3 and in ANNEX C for K=2 and K=4.The EM clustering algorithms result was generated by probabilistic descriptions of the clusters in terms of mean and standard deviation for the numeric attributes and value counts (incremented by 1 and modified with a small value to avoid zero probabilities) for the nominal ones.

Sine the results of the EM clustering method are probabilistic, every data point belongs to all clusters, but each assignment of a data point to a cluster has a different probability. Because the method allows for clusters to overlap, the sum of items in all the clusters may exceed the total items in the training set. This hold true for all EM out put.

Attributes	Values	Counts for each Value K=3 C0	C1	C2
SEX	Male	1567.26	2808.33	1240.40
	Female	3567.60	2179.75	476.64
MARSTAT	Divorced	892.70	140.36	72.92
	Married	1437.10	571.47	300.42
	Single	2153.75	4242.97	1302.26
	Other	16.37	29.78	5.83
	Widowed	637.92	6.48	38.58
EDUEXP	Primary	1289.18	749.04	8.77
	Sec_Old	2479.68	2249.04	224.26
	Sec_New	304.26	1441.92	181.81
	Tertiary	106.96	268.48	1292.54
	Other	957.75	282.59	12.64
OCCUPAT	Clerks_Sales	670.59	1037.44	132.96
	Other	2188.18	1102.77	71.03
	Machine_Craft	273.60	571.24	74.14
	Officials_Proff	89.90	116.50	1154.58
	Un_Empl	583.69	1176.79	262.51
	Unskilled_Occ	1318.04	940.59	3.36
	Armed_Force	15.83	47.72	23.44
REASHERE	Hard_Risk	3376.92	1918.97	876.10
	Planning	671.18	2084.69	478.12
	Marriage_Rel	168.18	431.37	105.44
	Sec_Test	696.75	309.48	194.75
	Other	224.81	246.56	65.62
EXPECT	Not_appl	46.57	34.11	13.30
	DonotKnow	3500.03	2783.02	938.93
	Negative	594.36	2116.57	627.06
	Positive	995.88	56.38	139.73
HIV	Negative	597.25	4378.92	1215.82
	Positive	4537.60	609.17	501.22
Tool Parameters	***Seed***	1000		
	Itertion	1000		
	Percentage	47%(1381)	38%(1145)	15%(433)
	prior Prob.	0.43	0.42	0.14
	Log value	-7.33		

Table 4. 12: EM cluster out put for K=3

The EM cluster out put displayed in the above tables (Table 4.12, and ANNEX C) were a point of discussion between the researcher and domain experts. From the discussion the following points are identified.

For K=4, as it is shown in (ANNEX C),there are four clusters (Co,C1,C2,and C3).From these four clusters C2 has small number of instances 13% (376).This figure is below the threshold value .If we drop this cluster ,it will be similar to cluster generated when K=3

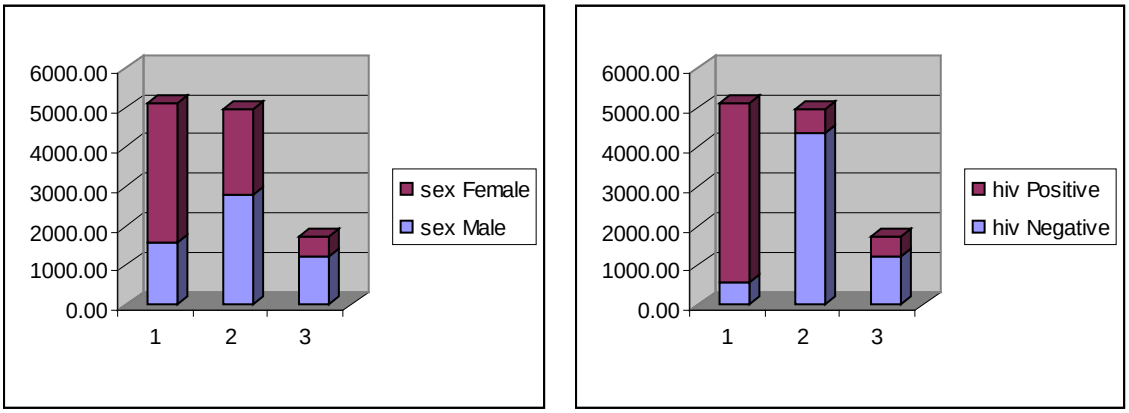
In comparing cluster formed by using K=2 and K=3, they are good cluster in terms of the percentage, prior probabilities, the variety of modal values they contain. Cluster C0, and C1 for both K=2 and K=3 are also similar characteristics in terms of modal attributes values. But cluster index C2 in K=3 contain clients which have some different values for educational background attribute and occupation title.

As it is stated in literature part of this research a good clustering are clusters that ensure the inter-cluster similarity is low and the intra-cluster similarity is high. Therefore the researcher ,along with the comments and suggestions of domain experts decided that that appropriate number of cluster were three (K=3),and this clustering model is taken for further analysis .

4.3.3.3 EM Cluster Interpretation

Interpretation of cluster obtained by using K=3, is presented as follows.

- Cluster index 0(C0): Nearly half (47%) of the total clients were clustered in this index. This cluster has the following features.
    - From the total of HIV positive clients 80% were under this cluster.
    - The majority (62%) of married clients are also in this category.
    - 80% of the divorced and 93% from the widowed clients are also in this cluster.
    - The 60% of the clients in this group visit the VCT centers because they think that they are in hard risk in relation to their serostatus and for confirmation of the previous result.
    - From the total number of clients who expects positive result in HIV, 83% are in this cluster.
    - Generally this cluster can be considered as clients who are in high risk in relation to HIV.



A **B**

Figure 4. 1: Sex (A) and HIV test result (B) distribution in Clusters

- Cluster index 1(C1): The second larges cluster which has 1145 (38%) instances is characterized by :
    - 70% of the HIV negative clients.
    - This cluster is dominated by male, single (unmarried), and unemployed clients (58% from the total unemployed).
    - Approximately 11% of the clients in this cluster are HIV positive. This cluster can be generalized by medium risk level.

- Cluster index 2(C2): One of the other three clusters which had the worst (the lowest count value for the attribute in any of the clusters) and contains 433 (15%) instances is described by:
    - The majority of male clients as compared with females,
    - 19 % of HIV negative.
    - The percentage of HIV positive clients is about 9% from the total HIV positive. This cluster can be generalized as low level of risk

Cluster Index	Level of risk	Priorities Assigned	Percentage	No. of Records
C0	High Risk	1	47%	1381
C1	Risky	2	38%	1145
C2	Low Risk	3	15%	433

Table 4. 13: summery of cluster description for k=3 in terms of risk

Once the cluster indexes are identified, classification model can be built easily. In the clustering sub phase, the main task was to identify possible clusters that can be considered as a dependant variable so that new instances can be classified accurately.

4.4 Classification sub phase

The major task here is to classify a data item into one of several predefined classes. The cluster model helps to segment the VCT clients successfully. But, classification helps to classify new clients in to different identified groups by the clustering phase. This classification problem can be solved by using decision tree and ANN. The following sub sections shows how the classification experiment is conducted and the comparison of classification based on cluster model and classification without cluster model.

For the classification experiment two decision variables are used. In the first part of this classification experiment HIV test result with negative and positive values is use a decision variable. On the second part of the experiment the decision variable used was the cluster index (C0, C1, and C2) which are represented by high risk, risky and low risk and obtained in the clustering experiment.

The total dataset (14793) record was used to construct the decision tree and ANN (Multilayer perceptron). These algorithms have been implemented using the WEKA and evaluated by 10 fold cross-validation test option.

4.4.1 Classification Experiment one

In the first classification experiment, all 14 attributes without the cluster index are used. To improve the accuracy and the performance of the algorithm attribute are selected by

using their information gain. The WEKA attribute selection out put result (HIV attribute as a target) using information gain is displayed in the following table

For the HIV attributes as a target class, the WEKA attribute selection algorithms rank attributes by using information gain. The selected attributes by their rank are MARSTAT, EXPECT, STIHIST, LASTSEX, EVERHAD, EDUEXP, SEXWORK, REASHERE, OCCUPAT, SEX, AGE, EMPLOYED, and CASUAL.

In the case of decision tree, two different kinds of algorithms J48 and Random Tree were tested. The result of the experiment and analysis was made first using all variables (14), second excluding, CASUAL and EMPLOYED, third CASUAL, EMPLOYED, and AGE and finally excluding CASUAL, EMPLOYED, AGE, SEXWORK, STIHIST, LASTSEX.

The result of the experiments is summarized on Table 4.14 below

Algorithm	No. of Attribute	No. of leaf nodes	Tree size	Time elapsed (sec)	Tree Type	Splitting Criteria	Accuracy %	Correctly classified instances
J48	14	105	151	1.28	Multi-way tree	Information Gain	74.96	11090
	12	96	133	0.75			74.9	11080
	11	66	87	0.36			74.21	10978
	7	49	65	0.23			72.09	10665
Random Tree	14		6588	0.49	Multi-way tree	Random	71.14	10525
	12		6169	0.47			71.16	10527
	11		5454	0.39			71.6	10593
	7		2286	0.27			71.49	10570

Table 4. 14: Analysis of decision tree algorithms result for HIV class

By analyzing the above table (table 4.14) and the confusion matrix of each classifier it was possible to compare models .When we see Random tree algorithm, the time required

to build the model is shorter when compared to J48.But the size of the tree is not manageable. Moreover the accuracy for each set of attribute is lower than J48.Therefore J48 performance is better than random tree.

4.4.2 Classification experiment two

Unlike the first experiment here the class or dependent variable used are cluster indexes which are the result of the cluster model. The result of the experiment using J48 and Random tree decision tree algorithm is presented and analyzed using the following table.

Parameters arranged: For J48, minpoints on the leaf node3, the confidence factors for pruning is .25. In the previous experiment it was set to 0.1. This is because smaller values will prune the tree and some important attribute may be eliminated. For the random tree the minpoint is also three.

Algorithm	No. of Attribute	No. of leaf nodes	Tree size	Time elapsed (sec)	Tree Type	Splitting Criteria	Accuracy %	Correctly classified instances
J48	15	121	158	0.13	Multi-way tree	Information Gain	94.39	2793
	13	122	159	0.05			94.35	2792
	12	116	152	0.03			94.25	2789
	8	113	145	0.06			94.35	2792
Random Tree	15		1559	0.13	Multi-way tree	Random	85.23	2522
	13		1522	0.08			85.9	2442
	12		1396	0.09			86.85	2570
	8		943	0.09			92.66	2742

Table 4. 15: Analysis of decision tree algorithms based on cluster index

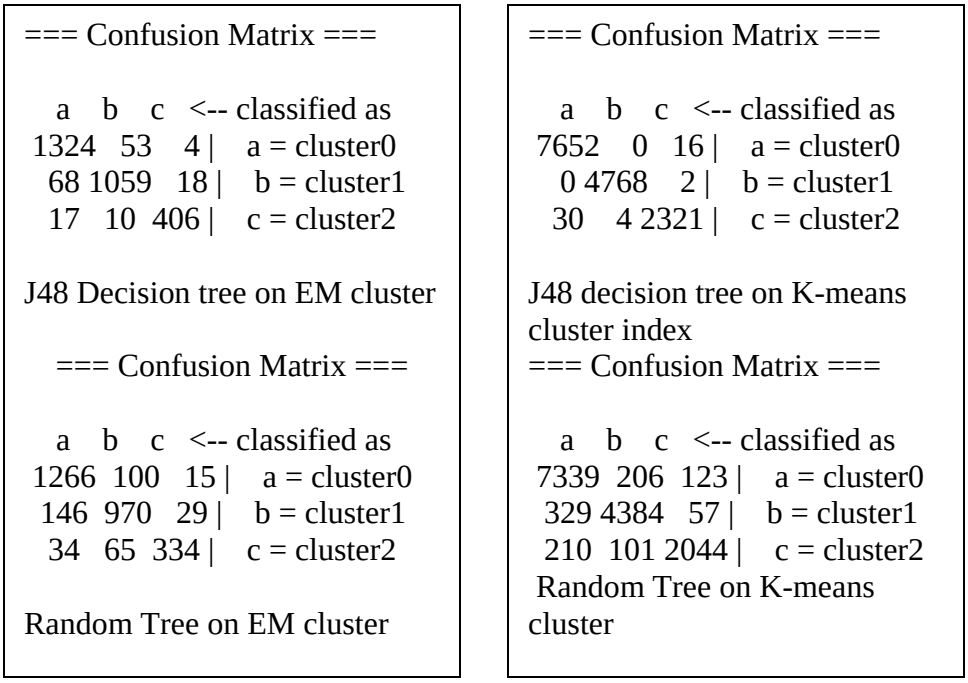


Figure 4. 2: Confusion Matrix of classifiers for EM and K-Means cluster index

Figure 4.2 shows the confusion matrix for the classification experiment (J48 and random tree) based on the results of the K-Means and EM cluster indexes. The confusion matrix shows the number of instances that are correctly and incorrectly classified.

4.4.3 Neural network Model Building experiment (ANN)

Another classification algorithm to build a model used in this research is artificial neural network. As it is stated in the literature part of this research ANN can be used as diagnosis tool in health organizations. To perform the experiment, the data set was preprocessed again so that it can be used by the selected WEKA tool. The algorithm that is supported by WEKA for this task is Mulitlayer Perceptron. ANN accepts inputs that

are in binary. The normalization process was handled by using the WEKA tool. All attributes except the target were normalized.

Once the data is preprocessed and ready, two experiments was performed. The first experiment was using HIV test result as target attribute and the second was using the cluster index found in the EM clustering experiment.

In the first experiment, similar to decision tree, different attribute set were tested. Different parameters was arranged to see the performance of the algorithms. The learning rate, number of hidden layers was modified and a result for which the performance is best was selected. Some of the experiment results are summarized in the following table.

N.of Attributes	Time elapsed(Sec)	No. of (nodes)Hidden Layers	Learning rate	Momentum	Accuracy	Instances classified
14	1012.06	9	0.3	0.2	74.69%	11050
14	679.64	6	0.3	0.2	75.46	11164
12	2385.51	Default	Default	default	74.63	11039
11	644.55	6	0.5	0.2	73.65	10896

Table 4. 16 : ANN for HIV class label

As it is shown in Table 4.16, the performance of the classifier for different variables and parameters was comparable to that of decision tree of J48 result. Here a slight improvement has been observed. The major problem is the time required to build the model. In terms of this the decision tree classifiers are fast.

In the second experiment the target attributes used was the cluster indexes of the K-means and the EM algorithms (K=3). This experiment was conducted in order to see the

performance of the algorithms on cluster index target. As it is done in the classification experiment the comparison was made by testing the performance of the classifiers using the cluster index. In this experiment the after a trial of different experiments for 12 attributes when the learning rate is 0.5,and the number of nodes in hidden layer is 6 the performance of the classifier were 98.68% for EM cluster index. That means the classifier correctly classifies 2920 in to the three cluster index. This is shown in the following Figure 4.3.

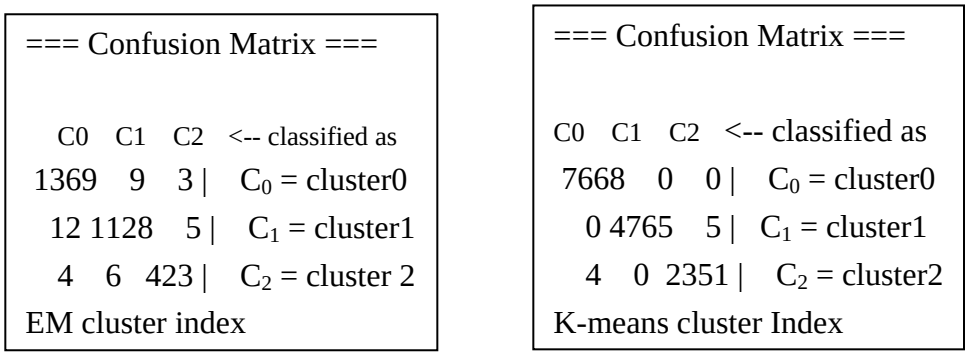


Figure 4. 3 : Confusion matrix of ANN on EM and K-Means cluster index

For the cluster index generated by the K-means the classifier parameters that result in better performance where when learning rate is 0.5.The classifier accuracy was 99.93%. From the confusion matrix we can se that the classifiers accuracy was 100% for cluster index (cluster 0) C0.

4.4.4 Comparison of models

In comparing the result of K-means and expectation maximization, as it is stated in literature and methodology of the research both algorithms has their own advantage and

disadvantage. The choice and efficiency of algorithms depend on the nature of the dataset. The comparison is made based on the accuracy of the decision tree and decision rule algorithms. Both algorithms are tested with K=3. The result of the classification algorithms using the cluster index as a class label for 12 attributes is displayed in the following table.

Algorithm	Dataset	Accuracy			No Instances
		Decision Tree			
		J48	Random Tree	ANN	
K-Means	Cluster 0	99.79	95.7	100%	7668
	Cluster 1	99.95	91.9	99.89%	4770
	Cluster 2	99.4	86.79	99.83	2355
EM	Cluster 0	96.23	90.87	99.13	1381
	Cluster 1	93.36	88.38	98.51	1145
	Cluster 2	93.07	83.14	97.69	433

Table 4. 17:10-fold cross validation accuracy for various algorithms

As indicated in Table 4.17, the K-means performs better than EM in both decision tree and decision rule algorithms. With respect to decision tree algorithms, J48 performs well for both K-means and EM cluster index. The performance of the neural network classifier is better than the decision tree classifier.

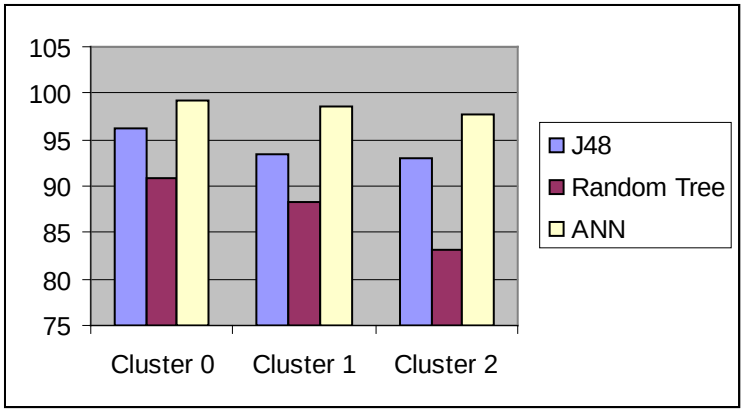


Figure 4. 4: Classifiers accuracy comparison

4.5 Discussion of Results

The results of experimentation on the dataset using Classification and Clustering algorithms as presented in the above tables and graphs were encouraging in giving reasonable knowledge and patterns from the dataset. The output for each test was described at the end of the tests as obtained from the experimentation. Based on such obtained results, the researcher tried to consult domain experts in counseling. This helps to evaluate the obtained results.

The following points, in addition to the discussion made during the experimentation in each phase of the experiment, are discussions on results based on the outputs and comments from experts.

Clustering sub phase: In doing the clustering experiment, after selecting the relevant features, the total dataset was tested to identify possible clusters. In this experiment 56468 records with 14 variables were tested. The result was not encouraging, because the

dataset was not balanced. To have a good model the algorithms should have a chance to learn about all kinds of records.

By using sampling techniques the dataset was made to be relatively proportional, the experiment was conducted by using both K-Means and EM algorithms. A number of clusters were identified .The selection of relevant clusters is made after a discussing with domain experts. In both algorithms the number of cluster K, selected was K=3.

In K-Means, the out put of the cluster algorithms was fair in terms of instances distribution in each cluster index C0, C1, and C2 (52%, 32%, and 16%). From the result it was found that:

- 84% of HIV positive clients were clustered in C0 and 90% of HIV negative clients in cluster index C1.
- In cluster index C0, Single Female clients whose educational background is secondary education in old curriculum were clustered. In addition 52% this cluster clients visit the VCT centers because they think that they are in hard risk with respect to HIV.

In EM 47%, 38%, and 15% instances were clustered in indexes C0, C1, and C2 respectively.

- 80% of HIV positive clients with the majority of Female (57%) are clustered in cluster index 0(C0).
    - What special pattern observed in this index was 93% of the widowed, 80% divorced and 62% of married clients are under this cluster.

- In K-means for index C0 (the majority of HIV positive), it was 10% for widowed, 14% for divorced and 24% for married.
- This cluster also includes 63% of clients whose educational background is primary, 50% secondary in old curriculum and 76% whose educational experience is grouped under “Other”. The value “Other” indicates clients who are able to read and illiterate.

- This cluster has got the attention of domain experts than the other because the instances behavior grouped in this cluster are expected to be positive in HIV.
- 70% of HIV negative with 50% male clients were under cluster index1 (C1).From this clients 10% of the clients are HIV positive.
    - 62% of clients whose occupation title are Machine_Craft (which includes Machine operators, crafts and trades, skilled agriculture and fishery) grouped in C1.
    - More than 74% of clients whose educational level is in secondary education by the new curriculum are grouped in this cluster.
- This cluster also provides information about group of clients who are at lower risk compared with cluster index 0.Even if the risk is considered as lower, counselors should treat this group of clients carefully.
- 15% of the total dataset are grouped in cluster index 2 (C2).This cluster is characterized by:
    - 77 % of tertiary levels clients are in this cluster.
    - Approximately 85% (from a total of officials and professionals) of clients are officials and professionals.

- 22% from a total of male clients and 16% from single are also in this cluster.

According to the experts evaluation on cluster made by K-means and EM, the clusters generated by EM were interesting than the K-means.

Classification sup phase: two classification model were built using two different class label (Target variable).In the first case the target variable used were the HIV test result of the clients and in the second case the cluster index (out put of cluster model). Decision tree and ANN algorithm were tested.

From the variety of decision tree algorithms J48 and Random tree were used. The J48 algorithm performance was better than the Random tree.

- For classification model using the HIV variable as target, the average performance of the J48 algorithm for four different input variables were 74.04%.
- Using the cluster model J48 performance for similar input variables were 94.33%.
- For Random tree it was 71, 34% for the first experiment and 87.6% for the second case.
- The over all classification model built using cluster model is by far better than the other model.

The ANN classifiers performances were the best result in this experiment. As it is indicated in literature part of the research, one of the main advantages of a neural network is its robustness to data noise. In all classification experiment conducted using the

clusters index as target variables; variables that were not used in the clustering phase were added. This variable was ignored in the clustering experiment. The main reason for their removal was most of this ignored attributes were skewed.

The assumption made at this point were, inclusion of more variables as possible will help the classifiers in to improve the accuracy in classifying new records to the cluster indexes. The neural network as it is mentioned above was not that much affected by these variables. That is why the accuracy was best as compared with the others.

In evaluating whether to use the K-means or EM for the classification task, the out put of the cluster model using K-means and EM were evaluated using the classifier accuracy.

- J48 performances were better on K-means cluster index than EM.
- J48 classifier classifies cluster index C0 of K-means algorithms with 99.79% accuracy while it was 96.23 % for EM.
- ANN performs well in K-means than EM
- The neural network classifier accuracy was better in K-means cluster index than EM indexes.
- All classifiers perform will on the K-means cluster index than EM cluster index.

Even if classifiers using the K-Means cluster index performs well the cluster model generated by EM was selected by domain experts. As it is stated in cluster sub phase discussion part above, the cluster index generated by EM was getting the attention of the domain experts. Due to this the EM cluster index are selected as an input for the classification model.

As it is stated in the methodology part of this research, one of the problems in relation to neural network is that the models, which are built by using neural network, are incomprehensible for a human and hence the extraction of business knowledge from it will be in question. In this regard the researcher concluded that J48 algorithm can be used to build the classification model.

CHAPTER FIVE

Conclusion & Recommendation

5.1 Conclusion

As stated in the first chapter of this research work, the underlying objective for undertaking this research was to explore the potential applicability of data mining techniques on VCT client dataset in developing a model that could support VCT, there by enabling policy makers, Health officers, Donors, NGO's, etc to make enhanced decisions in the effort to plan HIV/AIDS intervention program.

In order to meet the objective the researcher adopted a data mining techniques in the knowledge discovery process. In an entire process, a CRISP-DM process model was followed for the discovery of knowledge. The task of data mining here was clustering and classification. For analysis, exploratory data Analysis, which represents out puts by table and graph, method was used .The researcher evaluated and selected WEKA3.5.8 data mining tool to serve in the knowledge extraction and model building process.

The dataset for the research was VCT client records that were collected at OSSA counseling and testing centers which is supported by CDC. A total of 56468 records with 14 selected variables were used for the experimentation.

The goal to be achieved was to group VCT clients in to similar group based on their behavior in pre-test and post-test counseling session, and to build a classification model

that can label clients in one of the identified group. From the overall research the researcher concludes:

- VCT for HIV generates and stores a tremendous amount of data from pre-test, testing and post-test counseling session. These data should be analyzed and converted to valuable information so that it can supports decision making at different levels.
- Although data are collected and stored using the Epi-Info database in VCT centers that are supported by CDC like OSSA, the database is full of incorrect values, missing values and noisy fields. This has resulted in using only few numbers of fields for experimentation.
- In recording and filling the data that are collected from clients, there is no common understanding between counselors. This has created a problem during preprocessing.
- VCT centers collects large amount of data. But they don not use the information/data they have at hand. This is due the existing system doesn't support data mining functionalities like clustering and classification. For effective utilization of the data to support decision makers and data analyst, data mining techniques should be integrated with the existing database systems.
- From the clusters obtained it was possible to identify group of clients which are categorized in very high risk, high risk and low risk. Based on this
    - Gender, HIV test result, educational experience, martial status and occupation title are particularly important parameter in determining group membership.

- In terms of gender females are at very high risk when compared with males.
- In terms of martial status widowed, divorced, and married clients are at very high risk.
- With respect to educational experience clients who are illiterate and only able to read are at very high risk. Moreover students in secondary school are, especially those in old curriculum are at very high risk as compared with the others.
- Clients who are involved in unskilled occupation at a very high risk as compared with the others.

- Among from the data mining classification techniques examined, although both decision tree and neural network showed comparative accuracy and performance, the decision tree approaches seems more appropriate to the problem domain. This is because, the decision tree algorithms has a simple feature which can be easily understood by non technical staff.
- Although ANN classifier showed slightly better accuracy than decision tree classifier the researcher has concluded to use and build the classification model using decision tree. Tim's 2002 study (Shegaw, 2002) shows that, ANN have problems in health practice due to legacy ethics and scientific doubts. This is due to operations of the ANN as a black box. The adjustment of weights in back propagation method is not easily understood by non technical staff in VCT.
- Classification model built using the automatic cluster index is better than the model built by assigning class labels manfully.

- Although encouraging results are obtained in this research, due to the problem of processor speed, memory, and the requirement of uninterrupted power for more than two days in the case of clustering algorithm (for example EM), it was not possible to conduct experiment on variety of clustering and classification algorithms by changing parameters.

The results of the research were encouraging, as domain experts in the field accepted it, and proved that data mining can be applied in VCT so that stake holder in VCT can use the result of the research in the process of HIV intervention programs.

5.2. Recommendation

This research work shows the potential applicability of data mining techniques in VCT client dataset in identifying similar groups and developing a classification model. In this process, it was learnt that more research and development efforts need to be conducted to enable and explore the variety of data mining techniques that can be applied in VCT data. Based on the observations and experiment, the following recommendations can be put forwarded:

- For policy makers Health officers, Donors, NGO's:
  - Data related to HIV will have a great impact on the intervention program that will be designed globally. In this research much of the records found in Epi Info database were not used for experiment and analysis. Much of the time was spent in preprocessing. Data mining practices in such large organizations like CDC is very useful for good decision making and planning. Hence

storage mechanisms of historical data and data warehousing should be practiced.

- Continuous training and support should be given to counselors and data mangers in VCT who are directly involved in the counseling process.
- VCT center apart from reporting the records should be able to use the data collected at hand for their main duty. Hence integrating the existing system with data mining and expert system should be practiced.
- The clusters identified showed that females, widowed, divorced and married clients are at high risk. Moreover clients whose educational experience is at lower level (illiterate and able to read) are highly exposed to HIV. Students secondary education (specifically in old curriculum) levels are at higher risk than the others. Decision makers and planners should note and give attention to such variables in designing intervention programs.

- For further researches: In this research it was learnt that more research and development efforts need to be conducted to enable and explore the variety of data mining techniques that can be applied in VCT data.

  - The clustering and classification analysis proposed here are by no means exhaustive, only a limited number of factors were utilized. Future clustering and classification can be set up in a vast number of algorithms, depending on the ultimate goal.
  - For the clustering task, the employment of algorithms like CURE, that is good in handling categorical variable may yield better result. In addition

exploratory data analysis (EDA) can be used to identify systematic relations between variables when there are no complete priori expectations on variables.

- Although decision tree and ANN resulted in an encouraging out put, other techniques such as Belief network should be explored.
- A support vector machine (SVMs) which belong to a family of generalized linear models that achieves a classification or regression decision based on the value of the linear combination of features can be explored. SVMs possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy.
- The size of the dataset has an impact on data mining research. Especially proportional dataset will enhance the performance of algorithms. Further research can be conducted using large dataset by high speed processing machines.
- The dataset used for this research are from VCT centers. In this case, clients visit the centers purposefully. Dataset that are collected from mobile VCT centers and hot line telephone services of HIV counseling could be explored. This in turn would help to compare results obtained by using VCT dataset.
- The possibility of incorporating the findings of this study in another application should be explored.

Reference

1. Abreham Tesso, 2005, *The Application of Data Mining Technology to Identify Determinant Risk Factors of HIV Infection and To Find Their Association Rule*. Unpublished Master's Thesis. Addis Ababa University. Addis Ababa
2. Antenane Korra, Mebiratu Bejiga, Solomon Tesfaye, 2005, "Socio-demographic profile and prevalence of HIV infection among VCT clients in Addis Ababa", *Ethiop. J. Health Dev.*19(2) pp.110-115
3. Addis Ababa city Government HIV/AIDS Prevention and control office (AAHAPCO), 2007, *HIV/ AIDS Situation In Addis Ababa.* viewed 17 June 2008 Available at URL: http://www.aahapco.org/hivaidssituation.htm
4. *AIDS Epidemic Update, UNAIDS, December 2007.* viewed 16 June 2008. *Available at* http://www.etharc.org/aidscampaign/statistics.htm
5. AVERT, AVERTing HIV/AIDS2007 ,*Worldwide HIV & AIDS Statistics.* Viewed 15 June 2008. Available at URL:http://www.avert.org/worldstats.htm
6. Berry, M & Linoff G, 2004, Data mining techniques for marketing, sales and customer relationship management, 2nd edn, Wiley Publishing, Inc., Indianapolis, Indiana.
7. B. Leke Betechuoh, T. Marwala and T. Tettey ,2007, *Using Inverse Neural Networks for HIV Adaptive Control*
8. Business Intelligence to Marketing and Management (BI2M) service.
9. Buchanan, B.G. 2006, *Brief History of Artificial Intelligence* .Viewed 28 December 2008. Available at http://www.aaai.org/AITopics/bbhist.html

10. Centers for Diseases Control and Prevention (CDC), 2008. HIV/AIDS statistics and Surveillance. Viewed 20June 2008
Available at URL:http://www.cdc.gov/hiv/topics/surveillance/basic.htm#hivest

11. Cios, K.J., Pedrycz, W.,Kurgan ,L.A., and Swiniarski, R. ,2007, *Data Mining Methods a Knowledge Discovery Approach*, Springer Science +Business media, New York, U.S.A.

12. Corcoran, J. and J. A. Ware, 2001, Data *clustering using artificial neural networks as a precursor to crime Hot spot prediction. Viewed 11 Decemeber2008*
Available at URL: www.uqsrc.uq.edu.au/images/Jonathan_Corcoran.pdf

13. Dasu T. and Johnson T. (2003) "Exploratory Data Mining and Data Cleaning" Wiley Pub. Inc., Indianapolis: Indiana

14. David L. Olson and Dursun Delen, 2008, Advanced Data Mining Techniques Springer-Verlag Berlin Heidelberg

15. Denkew Abera (2003). The Application of Data Mining to support Customer relationship management at Ethiopian Airlines Unpublished Master's Thesis. Addis Ababa University. Addis Ababa

16. Dorian Pyle ,2003, *Business Modeling and Data Mining*, Morgan Kaufmann Publishers

17. Family Health International, 2005,*HIV Voluntary Counseling and Testing: Skills Training Curriculum Participant's Manual VCT TOOLKIT,*

18. Fayyad, U., Piatetsky-Shapiro, G & Smith, P. 1996, From Data Mining to Knowledge Discovery in Databases. Viewed 16 November 2008
Available at URL: http://citeseer.nj.nec.com/fayyad96from.html

19. Federal HIV/AIDS Prevention and Control Office Federal Ministry of Health, July 2007,*Guidelines for HIV Counseling and Testing in Ethiopia:* Addis Ababa

20. Hand, D., Mannila, H., Smytli, P. 2001,*Principles of Data Mining*, The MIT Press, Cambridge

21. Harleen Kaur and Siri Krishan Wasan ,2006, "Empirical Study on Applications of Data Mining Techniques in Healthcare", *Journal of Computer Science 2 (2),194-200*

22. Hiroto Saigo, Takeaki Uno , and Koji Tsuda ,2007**,** *Mining complex genotypic features for predicting HIV-1 drug resistance.* Viewed 15 June 2008 http://bioinformatics.oxfordjournals.org/misc/terms.shtml

23. I. Taniar, David, 2007, *Data mining and knowledge discovery technologies*, IGI publishing, Hershey , New York

24. IBM,2003b, *Decision Edge for Fraud and Abuse Management*. Viewed 10 September 2008. Available at http://www-3.ibm.com/software/data/bi/dccisionedge/de fam. htm

25. I. H.Witten and Eibe Frank, 2005, *Data mining practical machine learning Tools and Techniques,Morgan Kaufmann Publisher,U.S.A*

26. I.H Witen, and Frank, E. (2000). *Data Mining: practical machine learning tools and techniques with java implementations*. Morgan Kaufmann publishers. San-Francisco

27. Jain, A.K., Murty M.N., and Flynn P.J. 1999, *Data Clustering: A Review.* ACM Computing Surveys, Vol. 31, No. 3, pp.264-323

28. Kamber, M. and Han J. (2001). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann U.S.A

29. Kloos H and D. Haile Mariam, 2000,"HIV/AIDS in Ethiopia an Overview", *Northeast African Studies* , Vol. 7, No. 1, pp.13-40

30. Kloos, H . D. Haile Mariam, and Lindtjørn, B. ,2007, " The AIDS Epidemic in a Low-Income Country: Ethiopia" ,*Human Ecology Review, Vol. 14, No. pp. 1 39-55*

31. Kristin B. DeGruy, 2000,:Healthcare Applications of Knowledge Discovery in Databases by" *JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT*, vol. 14, no. 2, pp 59-69

32. Larose, T. D. 2005, DISCOVERING KNOWLEDGE IN DATA: An Introduction to Data Mining,

33. Larose,T.D.,2006,Data Mining Methods and Models, John Wiley & Sons, USA

34. Lori Bowen Ayre, 2006, *Data Mining for Information Professionals.* Viewed 17 July2008.
URL:http://techessence.info/files/Ayre_DataMiningForInformationProfessionals_June2006.pdf

35. M.S. Chen, J Han,P.S.Yu,1996, "Data Mining: An overview from a Database Perspective", *IEEE transactions on Knowledge and Data Engineering,*Vol.8.No 6 pp.866-883

36. Mehmed Kantardzic, 2003, *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons .

37. Michael L. Gargano and Bel G. Raggad ,1999, “Data mining – a powerful information creating tool”, *OCLC Systems & Services* Volume 15 · Number 2 · pp. 81–90

38. MayoClinic.Com, 2008, Infectious disease. Available at WWW.MayoClinic.com

39. National HIV/AIDS Council Secretariat (NACS). National Guideline for VCT in Ethiopia. October 2000.

40. Paquet, E.***, 2004***, *Exploring Anthropometric Data Through Cluster Analysis,* Digital Human Modeling for Design and Engineering (DHM), Oakland University, Rochester, Michigan, USA.

41. Pavel Berkhin,2002, *Survey of Clustering Data Mining Techniques,* Accrue Software Inc., San Jose, CA

42. Robert A. Nisbet ,2004, “How to Choose a Data Mining Suite”, DM Review Special Report**,** http://www.dmreview.com/specialreports/20040323/1000465-1.html

43. Shegaw Anagaw (2002). *Application of Data Mining Technology to Predict Child Mortality Patterns: The case of (Butajira Rural Health Project)BRHP*. Unpublished Master’s Thesis. Addis Ababa University. Addis Ababa

44. Sang Jun Lee and Keng Saiu ,2001, “A review of Data mining Techniques”, *Industrial Management and Data System*,101/1, pp 41-46

45. Stat Soft : at www.statsoft.com. Viewed 21 February 2009

46. Tibebe Betsha (2005). *The Application of Data Mining Technology to Identify Determinant Risk Factors Of HIV Infection and To Find Their Association Rule*. Unpublished Master’s Thesis. Addis Ababa University. Addis Ababa

47. Two Crows Corporation 1999, *Introduction to Data Mining and Knowledge Discovery*. Viewed 15 June 2008

Available at URL: http://www.twocrows.com/intro-dm.pdf

48. Two Crows Corporation, 2005**,** *Introduction to Data mining and Knowledge Discovery 3rd ed* . Viewed 15 December 2007. Available at URL: http://www.twocrows.com/intro-dm.pdf

49. UNAIDS,2007, *AIDS Epidemic Update* .

Available at http://www.etharc.org/aidscampaign/statistics.htm

50. UNAIDS, 2006, *Report on the Global IDS Epidemic: An UNAIDS 10th Anniversary Special Edition*. Geneva: UNAIDS and WHO.

51. UNAIDS,(2002),Voluntary Counseling and Testing (VCT) Technical update Viewed 20 June 2008.Available at URL:http://www.rhrc.org/resources/sti/hivaidsmanual

52. World Health Organization Priority interventions, 2008, *HIV/AIDS prevention, treatment and care in the health sector*

53. Wubitu Hailu, 2007, Resources *of HIV – Kulich Youth Reproductive Health and Development Organization (KYRHDO) Ethiopia.* Viewed 16 June 2008

Available at URL: http://www.hivandsrch.org/voices/gebrekristos.php

54. X. Wu , V. Kumar, J. Ghosh, Q. Yang,and H. Motoda ,2007, "Top 10 algorithms in data mining", *Knowl Inf Syst*

Appendices

ANNEX A

Ethiopian HIV Counseling and Testing Record

RECEPTION

Country	Region	Woreda	Site Code	Site type Circle one 1 Free standing (NGO based) 2 Mobile 3 Primary health care (health centre/ hospital) 4 Clinic 99 Other	Org type Circle one 1 NGO 2 Gov. 3 Private	Residence Circle one 1 Urban 2 Rural 9 Other

Client code Date of visit (dd/mm/yyyy)	Return visit Circle one 0 No 1 Yes	New client code Circle one 0 No 1 Yes 99 N/A	Age	Sex Circle one 1 Male 2 Female	Counsellor code Partner code *Couple code*

PRE-TEST COUNSELLING SESSION

Session type Circle one 1 Individual 2 Couple 3 Group 99 Other	Marital Status Circle one 1 Married 2 Never married 3 Separated 4 Divorced 5 Widowed 99 Other	Couple Type Circle one 1 Married 2 Premarital 3 Presexual 4 Sex partner 98 N/A 99 Other	Education Circle one 0 Illiterate 1 Able to read 2 Primary 3 Secondary 4 Tertiary 99 Other	Employed Circle one 0 No (Inactive) 1 Yes (Active)	Occupation Circle one 1 Legislators, Sr. Officials, Managers 2 Professionals 3 Technicians, Ass. Professionals 4 Clerks 5 Service, Shop, Market, Sales	6 Skilled Ag. & Fishery workers 7 Crafts & trades 8 Plant/Machine Op., Assembly 9 Elementary occupation 10 Students 99 Other

Heard of the Service:	Client referred by:	Primary reason here		Suspected expos. time
Circle all that apply 1 Radio 2 Outreach 3 Posters 4 Other clients 5. Newspaper/magazine 6 Health institution 7 Telephone hotline 8 Anti-AIDS clubs 9 A PLWHA 10 Friends and family 11 CBO 12 TV 98 N/A 99 Other	Circle one. 1 Self – not referred 2 Public health institution 3 Private health institution 4 Military health institution 5 Friend or relative 6 Community-based organisation 7 NGO 8 School 9 Religious institution 10 Client 11 A PLWHA 99 Other	Circle one 1 Client risky/Had risk 2 Partner risky/Had risk 3 Doesn't trust partner 4 III/Symptoms 5 Premarital 6 Marital reunion 7 Family planning 8 Visa applicant 9 Referred 10 2nd Test (win.) 11 Confirm positive result 12 Get results of prev. test 13 Need counselling 14 Test before pregnant	15 Pregnant, must know 16 Plan for future 17 Death/illness of partner 18 Occupational exposure 19 Other blood/fluid exp. 20 Sexual assault 99 Other	Circle one 1 <1 month 2 1 to 3 months 3 4 to 6 months 4 Over 6 months 98 N/A

Previously tested?	Date prev. test?	Where prev. tested?	Ever had sex?	Condom use last 3 mos.	Used condom last sex?	History of STI
Circle one 0 No 1 Yes, HIV+ 2 Yes, HIV- 3 Yes, inconclusive 4 Result not given 5 Didn't take results 99 Other	___ ___ Month Year	Circle one 1 NGO 2 Public health Inst. 3 Private Health Inst. 98 N/A 99 Other	Circle one 0 No 1 Yes	Circle one 0 Never 1 Always 2 Sometimes 98 N/A	Circle one 0 No 1 Yes 97 Doesn't remember 98 N/A	Circle one 0 No 1 Yes 97 Don't know

No. Casual partners last 6 mos. No. Steady partners last 6 mos.	Is client sex worker?	Is client pregnant?	Result client expects	Pre-test partner notification plan
	Circle one 0 No 1 Yes 2 Counsellor thinks so 97 don't know 98 N/A	Circle one 0 No 1 Yes 97 Don't know 98 N/A	Circle one 0 Negative 1 Positive 97 Don't know 98 N/A	Circle one 0 Refused to notify 1 Agree to notify 2 Plan to notify 3 Unsure 98 N/A

POST-TEST COUNSELLING SESSION

Couple discordant
Circle one
0 No
1 Yes
98 N/A

Refused results
Circle one
0 No
1 Yes
98 N/A

Condoms accepted
Circle one
0 No
1 Yes
98 None available

Number condoms given
_____

Risk reduction plan developed
Circle one
0 No
1 Yes

Post-test partner notification plan
Circle one
0 Refused to notify
1 Agree to notify
2 Plan to notify
3 Unsure
98 N/A

Client referred to:
Circle all that apply
1 Follow-up counselling
2 Social services
3 TB clinic
4 Hospital
5 A PLWHA
6 STD Clinic
7 Family planning
8 Post–test club
98 N/A
99 Other

SERVICES RENDERED

	Circle one
Refused services	0 No 1 Yes
Counselled	Circle one 0 No 1 Yes
Gave test	Circle one 0 No 1 Yes
Received results	Circle one 0 No 1 Yes
Referred	Circle one 0 No 1 Yes
Condom demonstration	Circle one 0 No 1 Yes

Other: _____

Comments: _____

_____

_____

Counsellor's Signature: _____

ANNEX B

J48 algorithm out put

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 3
Relation: VCT Client data(14793)_clustered-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R11-
weka.filters.unsupervised.attribute.Remove-R5-
weka.filters.unsupervised.attribute.Remove-R1
Instances: 2959
Attributes: 12
SEX
MARSTAT
EDUEXP
OCCUPAT
REASHERE
EVERHAD
LASTSEX
STIHIST
SEXWORK
EXPECT
HIV
Cluster
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

HIV = Negative
| EDUEXP = Sec_Old
| | OCCUPAT = Machine_Craft: cluster1 (54.0/1.0)
| | OCCUPAT = Officials_Proff: cluster2 (51.0/2.0)
| | OCCUPAT = Clerk_Sales: cluster1 (124.0/5.0)
| | OCCUPAT = Unskilled_occ
| | | EXPECT = Negative: cluster1 (31.0)
| | | EXPECT = Positive: cluster0 (6.0/1.0)
| | | EXPECT = DonotKnow: cluster1 (61.0/4.0)
| | | EXPECT = Not_appl: cluster1 (1.0)
| | OCCUPAT = Other
| | | MARSTAT = Single

```
|  |  |  |  EXPECT = Negative: cluster1 (56.0)
|  |  |  |  EXPECT = Positive: cluster0 (5.0)
|  |  |  |  EXPECT = DonotKnow: cluster1 (69.0)
|  |  |  |  EXPECT = Not_appl: cluster1 (0.0)
|  |  |  MARSTAT = Married: cluster1 (18.0/3.0)
|  |  |  MARSTAT = Widowed: cluster0 (4.0)
|  |  |  MARSTAT = Divorced
|  |  |  |  SEXWORK = Not_appl: cluster1 (3.0/1.0)
|  |  |  |  SEXWORK = No: cluster0 (6.0/1.0)
|  |  |  |  SEXWORK = Yes: cluster0 (0.0)
|  |  |  MARSTAT = Other: cluster1 (0.0)
|  |  OCCUPAT = Un_Empl: cluster1 (70.0)
|  |  OCCUPAT = Armed_Force: cluster1 (6.0/1.0)
|  EDUEXP = Tertiary
|  |  OCCUPAT = Machine_Craft: cluster2 (16.0/4.0)
|  |  OCCUPAT = Officials_Proff: cluster2 (132.0)
|  |  OCCUPAT = Clerk_Sales
|  |  |  EXPECT = Negative: cluster1 (13.0/2.0)
|  |  |  EXPECT = Positive: cluster2 (0.0)
|  |  |  EXPECT = DonotKnow
|  |  |  |  EVERHAD = Yes: cluster2 (11.0)
|  |  |  |  EVERHAD = No: cluster1 (3.0/1.0)
|  |  |  EXPECT = Not_appl: cluster2 (0.0)
|  |  OCCUPAT = Unskilled_occ: cluster1 (9.0)
|  |  OCCUPAT = Other
|  |  |  SEX = Male
|  |  |  |  REASHERE = Planning: cluster1 (1.0)
|  |  |  |  REASHERE = Hard_Risk: cluster2 (11.0/1.0)
|  |  |  |  REASHERE = Marriagr_Rel: cluster1 (4.0)
|  |  |  |  REASHERE = Sec_Test: cluster2 (2.0)
|  |  |  |  REASHERE = Other: cluster2 (0.0)
|  |  |  SEX = Female: cluster1 (6.0)
|  |  OCCUPAT = Un_Empl
|  |  |  SEX = Male: cluster2 (39.0)
|  |  |  SEX = Female
|  |  |  |  REASHERE = Planning: cluster1 (11.0/1.0)
|  |  |  |  REASHERE = Hard_Risk: cluster2 (18.0)
|  |  |  |  REASHERE = Marriagr_Rel: cluster1 (2.0/1.0)
|  |  |  |  REASHERE = Sec_Test: cluster2 (1.0)
|  |  |  |  REASHERE = Other: cluster1 (1.0)
|  |  OCCUPAT = Armed_Force: cluster2 (1.0)
|  EDUEXP = Primary
|  |  MARSTAT = Single: cluster1 (151.0/2.0)
|  |  MARSTAT = Married
|  |  |  SEX = Male: cluster1 (14.0)
|  |  |  SEX = Female
```

```
|  |  |  |  EXPECT = Negative: cluster1 (9.0)
|  |  |  |  EXPECT = Positive: cluster0 (1.0)
|  |  |  |  EXPECT = DonotKnow
|  |  |  |  |  REASHERE = Planning: cluster1 (4.0)
|  |  |  |  |  REASHERE = Hard_Risk: cluster0 (6.0)
|  |  |  |  |  REASHERE = Marriagr_Rel: cluster0 (0.0)
|  |  |  |  |  REASHERE = Sec_Test: cluster0 (2.0)
|  |  |  |  |  REASHERE = Other: cluster1 (3.0/1.0)
|  |  |  |  EXPECT = Not_appl: cluster1 (0.0)
|  |  MARSTAT = Widowed: cluster0 (5.0)
|  |  MARSTAT = Divorced
|  |  |  SEX = Male: cluster1 (5.0/2.0)
|  |  |  SEX = Female: cluster0 (14.0/1.0)
|  |  MARSTAT = Other: cluster1 (1.0)
|  EDUEXP = Other
|  |  EXPECT = Negative: cluster1 (35.0/2.0)
|  |  EXPECT = Positive: cluster0 (5.0/1.0)
|  |  EXPECT = DonotKnow
|  |  |  OCCUPAT = Machine_Craft: cluster1 (5.0)
|  |  |  OCCUPAT = Officials_Proff: cluster2 (4.0/1.0)
|  |  |  OCCUPAT = Clerk_Sales: cluster1 (6.0)
|  |  |  OCCUPAT = Unskilled_occ: cluster0 (28.0/7.0)
|  |  |  OCCUPAT = Other
|  |  |  |  EVERHAD = Yes: cluster0 (11.0/1.0)
|  |  |  |  EVERHAD = No: cluster1 (6.0/1.0)
|  |  |  OCCUPAT = Un_Empl
|  |  |  |  REASHERE = Planning: cluster1 (3.0)
|  |  |  |  REASHERE = Hard_Risk: cluster0 (6.0)
|  |  |  |  REASHERE = Marriagr_Rel: cluster0 (0.0)
|  |  |  |  REASHERE = Sec_Test: cluster0 (0.0)
|  |  |  |  REASHERE = Other: cluster0 (0.0)
|  |  |  OCCUPAT = Armed_Force: cluster0 (0.0)
|  |  EXPECT = Not_appl: cluster0 (2.0)
|  EDUEXP = Sec_New
|  |  OCCUPAT = Machine_Craft: cluster1 (21.0/1.0)
|  |  OCCUPAT = Officials_Proff
|  |  |  SEX = Male: cluster2 (21.0)
|  |  |  SEX = Female: cluster1 (4.0/1.0)
|  |  OCCUPAT = Clerk_Sales: cluster1 (52.0)
|  |  OCCUPAT = Unskilled_occ: cluster1 (36.0)
|  |  OCCUPAT = Other: cluster1 (49.0/1.0)
|  |  OCCUPAT = Un_Empl: cluster1 (134.0/1.0)
|  |  OCCUPAT = Armed_Force: cluster1 (4.0)
HIV = Positive
|  EDUEXP = Sec_Old
|  |  EXPECT = Negative
```

```
|  |  |  REASHERE = Planning
|  |  |  |  OCCUPAT = Machine_Craft: cluster1 (4.0/1.0)
|  |  |  |  OCCUPAT = Officials_Proff: cluster2 (3.0)
|  |  |  |  OCCUPAT = Clerk_Sales: cluster1 (4.0)
|  |  |  |  OCCUPAT = Unskilled_occ: cluster1 (6.0/2.0)
|  |  |  |  OCCUPAT = Other
|  |  |  |  |  SEX = Male: cluster1 (3.0)
|  |  |  |  |  SEX = Female: cluster0 (5.0)
|  |  |  |  OCCUPAT = Un_Empl: cluster1 (1.0)
|  |  |  |  OCCUPAT = Armed_Force: cluster1 (0.0)
|  |  |  REASHERE = Hard_Risk: cluster0 (51.0/2.0)
|  |  |  REASHERE = Marriagr_Rel: cluster1 (6.0/2.0)
|  |  |  REASHERE = Sec_Test: cluster0 (12.0/1.0)
|  |  |  REASHERE = Other: cluster0 (4.0/2.0)
|  |  EXPECT = Positive: cluster0 (113.0/1.0)
|  |  EXPECT = DonotKnow: cluster0 (451.0/12.0)
|  |  EXPECT = Not_appl: cluster0 (5.0)
|  EDUEXP = Tertiary
|  |  OCCUPAT = Machine_Craft: cluster2 (5.0)
|  |  OCCUPAT = Officials_Proff: cluster2 (82.0)
|  |  OCCUPAT = Clerk_Sales: cluster2 (6.0/2.0)
|  |  OCCUPAT = Unskilled_occ: cluster0 (1.0)
|  |  OCCUPAT = Other: cluster0 (13.0/2.0)
|  |  OCCUPAT = Un_Empl: cluster2 (12.0/1.0)
|  |  OCCUPAT = Armed_Force: cluster2 (1.0)
|  EDUEXP = Primary: cluster0 (354.0/4.0)
|  EDUEXP = Other: cluster0 (199.0/3.0)
|  EDUEXP = Sec_New
|  |  OCCUPAT = Machine_Craft
|  |  |  SEXWORK = Not_appl: cluster1 (7.0/3.0)
|  |  |  SEXWORK = No: cluster0 (3.0/1.0)
|  |  |  SEXWORK = Yes: cluster0 (0.0)
|  |  OCCUPAT = Officials_Proff: cluster2 (8.0)
|  |  OCCUPAT = Clerk_Sales
|  |  |  SEX = Male: cluster1 (10.0/2.0)
|  |  |  SEX = Female: cluster0 (13.0/2.0)
|  |  OCCUPAT = Unskilled_occ
|  |  |  EVERHAD = Yes: cluster0 (15.0/1.0)
|  |  |  EVERHAD = No: cluster1 (4.0/1.0)
|  |  OCCUPAT = Other: cluster0 (30.0/4.0)
|  |  OCCUPAT = Un_Empl
|  |  |  REASHERE = Planning: cluster1 (9.0)
|  |  |  REASHERE = Hard_Risk: cluster0 (21.0/5.0)
|  |  |  REASHERE = Marriagr_Rel: cluster0 (0.0)
|  |  |  REASHERE = Sec_Test: cluster0 (3.0)
|  |  |  REASHERE = Other: cluster0 (0.0)
```

| | OCCUPAT = Armed_Force: cluster0 (2.0/1.0)

Number of Leaves : 116

Size of the tree : 152

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

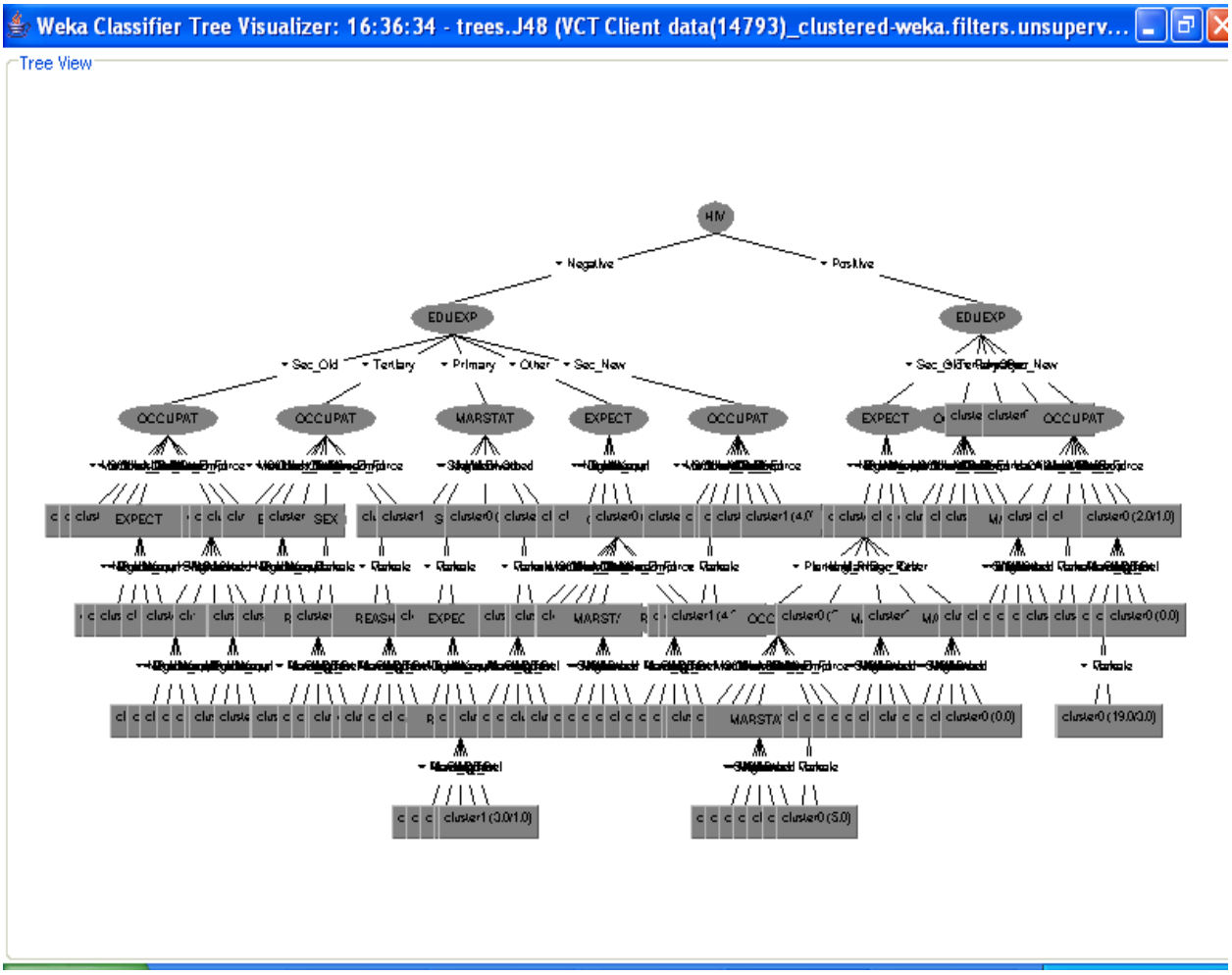
Correctly Classified Instances	2789	94.2548 %
Incorrectly Classified Instances	170	5.7452 %
Kappa statistic	0.9058	
Mean absolute error	0.0534	
Root mean squared error	0.1814	
Relative absolute error	13.1188 %	
Root relative squared error	40.1844 %	
Total Number of Instances	2959	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.959	0.054	0.94	0.959	0.949	0.971	cluster0
0.925	0.035	0.944	0.925	0.934	0.966	cluster1
0.938	0.009	0.949	0.938	0.943	0.977	cluster2

=== Confusion Matrix ===

a	b	c	<-- classified as
1324	53	4	a = cluster0
68	1059	18	b = cluster1
17	10	406	c = cluster2


Weka Classifier Tree Visualizer: 16:36:34 - trees.J48 (VCT Client data(14793)_clustered-weka.filters.unsuperv...
Tree View
HIV
Negative
Positive
EDUEXP
EDUEXP
Sec_Old
Tertiary
Primary
Other
Sec_New
OCCUPAT
OCCUPAT
MARSTAT
EXPECT
OCCUPAT
EXPECT
OCCUPAT
EXPECT
SEX
REASH
EXPEC
MARSTA
cluster0 (2.0/1.0)
cluster0 (0.0)
cluster0 (19.0/3.0)
cluster1 (3.0/1.0)
cluster0 (5.0)

ANNEX C

EM cluster partial view for K=2 and K=4

Attributes	Values	Number of clustered Instances(counts)	
		K=2	
		Cluster 0	Cluster 1
SEX	Male	1673.72	3941.27
	Female	3766.88	2456.11
MARSTAT	Divorced	937.16	167.83
	Married	1529.11	778
	Single	2298.4	5399.59
	Other	17.15	33.84
	Widowed	661.76	20
EDUEXP	Primary	1337.73	708.26
	Sec_Old	2558.42	2393.57
	Sec_New	313.72	1613.27
	Tertiary	232.66	1434.33
	Other	1001.66	250.93
OCCUPAT	Clerks_Sales	693.52	1146.47
	Other	2262.10	1098.89
	Machine_Craft270.	270.29	647.70
	Officials_Proff	203.70	1156.29
	Un_Empl	610.99	1411.00
	Unskilled_Occ	1384.63	876.36
	Armed_Force	20.36	65.63
REASHERE	Hard_Risk	3544.92	2626.07
	Planning	727.41	2505.58
	Marriage_Rel	190.25	513.74
	Sec_Test	733.52	466.47
	Other	247.49	288.50
EXPECT	Not_appl	47.75	45.24
	DonotKnow	3700.89	3520.10
	Negative	649.29	2687.70
	Positive	1044.67	146.32
HIV	Negative	768.41	5422.58
	positive	4672.19	974.80

Tool Parameters	Seed	1000	
	Iteration	1000	
	Percentage	49%(1442)	51%(1517)
	prior Prob.	0.46	0.54
	Log value	-7.44	

Attributes	Values	Counts for each Value K=4 Co	C1	C2	C3
	Male	895.86	1079.57	1152.69	2488.86
SEX	Female	3690.65	1852.06	394.80	287.46
	Divorced	854.43	51.24	69.48	131.82
	Married	1248.98	237.38	275.11	548.51
	Single	1847.34	2618.34	1162.81	2071.49
	Other	16.75	24.10	5.32	6.81
MARSTAT	Widowed	622.01	3.56	37.74	20.66
	Primary	1152.36	324.39	2.82	568.41
	Sec_Old	2100.07	975.60	133.52	1744.79
	Sec_New	297.18	1165.35	155.61	310.83
	Tertiary	99.56	281.03	1247.79	40.59
EDUEXP	Other	940.32	188.24	10.74	114.68
	Clerks_Sales	497.66	386.23	125.92	832.17
	Other	2108.77	712.62	70.89	470.69
	Machine_Craft	80.70	69.93	71.27	698.08
	Officials_Proff	72.02	84.08	1080.80	125.09
	Un_Empl	641.66	1199.12	180.39	2.80
	Unskilled_Occ	1184.38	468.34	3.78	606.48
OCCUPAT	Armed_Force	6.29	16.28	19.42	45.98
	Hard_Risk	2963.32	913.05	805.71	1490.89
	Planning	620.75	1419.15	409.34	785.74
	Marriage_Rel	153.86	238.40	94.42	219.30
	Sec_Test	627.28	162.62	180.82	231.27
REASHERE	Other	224.29	201.40	60.18	52.11
	Not_appl	41.77	17.74	11.58	23.89
	DonotKnow	3118.82	1444.96	845.29	1813.90
	Negative	554.96	1463.37	559.05	761.60
EXPECT	Positive	872.95	7.55	133.56	178.92
	Negative	576.84	2732.35	1086.72	1797.07
HIV	positive	4009.68	199.28	460.77	979.24
	Seed	1000.00			
	Itertion	1000.00			
	Percentage	42%(1254)	23%(674)	13%(376)	22%(655)
	Prior prob.	0.39	0.25	0.13	0.23
Tool Parameters	***Log value***	-7.30			

Declaration

I, the undersigned, declare that this 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.

_____

Biru Asmare

January 2009

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

_____

Ermias Abebe