ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING FOR EFFECTIVE PROGNOSIS OF HIV/AIDS TESTING

BERHANU FETENE

October, 2016

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING FOR EFFECTIVE PROGNOSIS OF HIV/AIDS TESTING

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

By

BERHANU FETENE

October, 2016

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING FOR EFFECTIVE PROGNOSIS OF HIV/AIDS TESTING

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

By

BERHANU FETENE

Name and signature of advisor and members of the examining board

Name		Signature
_____	Advisor	_____
_____		_____
_____		_____
_____		_____

DECLARATION

I, declare that this thesis is my own work except to the extent indicated in the acknowledgements and the references. It is being submitted for the degree masters of Science in Information Science at the University Addis Ababa Ethiopia. It has not been submitted before for any degree or examination at this or any other University.

_____

Student's signature

ACKNOWLEDGMENT

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

I would like to express my gratitude and heartfelt thanks to my advisor, Dr. Getachew H/Mariam for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study. His interest and encouragement has always stimulated me to accelerate to the completion of the work.

I would also thank my family who has been always with me through this study by encouraging and asking me about the progress of my thesis work.

Finally I would also like to thank sincerely all my friends those who helped me with their valuable support during the entire process of this thesis.

List of Figures

LIST OF TABLES

LIST OF ABBREVIATIONS

CAD…….Computer Assisted Design

CDC……..Centers for Disease Control and Prevention

CSA……Central Statistical Agency

DBMS…….database management systems

DFID……...United Kingdom for International Development

DHS……..Demographic and Health Surveys

HAPCO…….HIV/AIDS Prevention and Control Office

HIV/AIDS…….human immunodeficiency virus/ acquired immunodeficiency syndrome

ICF…….. ……International Coach Federation

KDD……Knowledge Discovery in Databases

MLP……. Multilayer Perceptron

SMOTE…….Synthetic Minority Oversampling Technique

SPSS…….. …….statistical packages for the social sciences

UNAIDS……..United Nations Program on HIV/AIDS

UNFPA……United Nations Population Fund

UNICEF……..United Nations Children's Fund

USAID……United States Agency for International Development

VRML……Virtual Machine Language

WEKA……Waikato Environment for Knowledge Analysis

WHO…….World Health Organization

Table of Contents

ABSTRACT

AIDS is the disease caused by HIV, which weakens the body's immune system until it can no longer fight off the simple infections that most healthy people's immune system can resist. HIV/AIDS continues to be a major global health priority. Knowing about HIV/AIDS help the individuals, communities, government and stakeholders to prevent and design the appropriate policy which leads to eradicate the disease.

Now days Healthcare and non-Healthcare industries are stored huge amounts of data about patients, hospital resources, disease diagnosis, electronic patient records, and medical devices. Huge amount of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost savings and decision making. The problem of effectively utilizing this huge amount of data is becoming a major problem for all health and to make effective decision. So in order to make effective and efficient decision on the data using data mining techniques is the best solution.

The main objective of this study was to design/develop an effective HIV/AIDS prognosis model by using data mining techniques.

The six steps hybrid methodology has been followed to develop the HIV testing prognosis model. Three classifications techniques such as Decision tree J48, Naivebayes and Neural network (MLP) algorithms were experimented for building and evaluating the models.

A total of 14,786 EDHS 2011 records have been used to develop the models. , the J48 pruned tree classifier algorithm with 10-fold cross validation test option performed best classification accuracy of 98.5121% with 98.51% sensitivity, where as Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option performed best classification specificity of 98.1% and 98.1% result respectively.

In general the results obtained from this research indicate that data mining is useful and appropriate tool in bringing relevant information for decision and policy makers. Hence, the organization should have to design a knowledge base system, which can provide advice for the domain experts is future research direction of the study.

Chapter one
Introduction

1.1. Background

The Healthcare industry has experienced a proliferation of innovations aimed at enhancing life expectancy, quality of life, diagnostic and treatment options, as well as the efficiency and cost effectiveness of the Healthcare system (Vincent et al, 2010), cited in (Asia , 2012). Moreover, primary health care plays a central role in health care systems worldwide. It can offer families cost effective services close to their home. Particularly in developing countries community health centers usually offer a broad range of services, including prenatal care, immunizations, treatment of childhood illnesses, treatment of malaria, tuberculosis and other common infectious diseases, and other basic medical care.

The main health problem of Ethiopia's are communicable deceases occurred by poor hygiene and improper nutrition. This situation is further aggravated by the high population growth, and high turnover in the number of physicians, pharmacists and pharmacy technicians. Hence health policy emphasizes the importance of achieving access, for all segments of the population, to a basic package of quality primary health care services. This service package should include preventive, promotive and basic curative services. Basically, design a strong policy that focuses on transmission diseases, namely HIV/AIDS, TB/Tuberculosis, Malaria; and maternal and childcare health problems. The policy integrated Health Sector Development Program in response to prevailing and newly emerging health problems in Ethiopia and in recognition of weaknesses in the existing health delivery system and Disseminate Information on health, hygiene and nutrition to inform the people (MOH).

HIV/AIDS has already become one of the greatest challenges of humanity. Since its emergence more than two decades ago, HIV/AIDS has victimized tens of millions of people from around the globe (WHO, 2007).

HIV (human immunodeficiency virus) is a virus that attacks the immune system, the body's natural defense system. Without a strong immune system, the body has trouble fighting off disease. Both the virus and the infection it causes are called HIV. White blood cells are an important part of the immune system. HIV infects and destroys certain white blood cells called CD4+ cells. If too many CD4+ cells are destroyed, the body can no longer defend itself against infection.

The last stage of HIV infection is AIDS (acquired immunodeficiency syndrome). People with AIDS have a low number of CD4+ cells and get infections or cancers that rarely occur in healthy people. These can be deadly. But having HIV doesn't mean you have AIDS. Even if, HIV can be treated in the health care's in well-organized way, but it has a problem on the diagnosis of the disease. Because HIV takes a long time without treatment to progress to AIDS usually 10 to 12 years and it occurs with co infection diseases such as Malaria, respiratory tract infections,

tuberculosis (TB), sexually transmitted infections (STIs), skin infections, and hepatitis. This makes the disease difficult to diagnosis and attack the immune system of the body natural defense system (WHO, 2007).

HIV/AIDS is a major public health concern and cause of death in many parts of Africa. Countries in North Africa and the Horn of Africa have significantly lower prevalence rates, as their populations typically engage in fewer high risks cultural patterns that have been implicated in the virus's spread in Sub Saharan Africa. Southern Africa is the worst affected region on the continent. HIV was first detected in Ethiopia in 1984 and the first two AIDS cases were reported in 1986 (MOH). There are many factors that promote the spread of the HIV in Ethiopia. The major factors are 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 mother to her fetus/child during pregnancy and breast feeding.

1.2. Statement of the problem

Healthcare's are sectors or organizations within the health system that provides a goods and services to treat and diagnosis the patients with curative, preventive, rehabilitative, and palliative care. In Ethiopia there are a lot of health care that treat and diagnosis many diseases. HIV/AIDS is a great problem for most developing countries because of the low diagnosis and treatment opportunities (WHO, 2007). The underlying research problem that initiated to do this research is the existence of high death rate of patients in HIV and AIDS in Ethiopian referral and zonal hospital those provides comprehensive HIV/AIDS care and treatment services. Even if the hospitals carried out a good and periodic treatment for its patients, there is a problem on the prognosis of the disease. As described by the experts of the hospitals, the late prognosis and diagnosis of HIV infection associated with increased morbidity, mortality, and probability of transmission. So the problem on the late diagnosis of the disease leads to delay the eradication of the disease and death of lots of patients.

Whereas, in the Healthcare organizations there are lots of data stored in manual formats, However, due to lack of computerized databases and useful analysis tools to realize hidden relationships and trends in data used to solve the problems faced by Healthcare professionals, planners and policy makers to identify the major factors for effective diagnosis of HIV/AIDS and in order to plan and implement effective HIV/AIDS control program to decrease the death of patients in Ethiopia.

HIV/AIDS is the major problem in developing countries that attacks mainly the major workforces of the countries and it continues as the major health problem in these countries (WHO, 2007). The major problem that makes HIV/AIDS diagnosis more severe is, since it occurs with co infection diseases such as Malaria, respiratory tract infections, tuberculosis (TB), sexually transmitted infections (STIs), skin infections, and hepatitis.

On the other hand, Healthcare industry today stored huge amounts of data about patients, hospital resources, disease diagnosis, electronic patient records, and medical devices. Huge amount of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost savings and decision making. The problem of effectively utilizing this huge amount of data is becoming a major problem for all health to make effective decision.

Additionally, inaccurate data leads to wrong decision. So in order to make effective and efficient decision it needs a new and more effective technique to extract the hidden knowledge that leads the user to the right decision. This has lead to the exploration of a new field of research called data mining. Data mining refers to computer aided pattern discovery of previously unknown interrelationships and recurrences across seemingly unrelated attributes in order to predict actions, behaviors and outcomes (Madan, 2006). Data mining helps to identify patterns and relationships in the data.

However, lack of data in the database is the main problem of developing countries like Ethiopia. In Ethiopia, the practical challenge for health care professionals and stakeholders working in the health care's are lack of timely and reliable health information on the health states of defined population groups.

Different researches have been done on HIV/AIDS prognosis in Ethiopia and abroad, such as Tesfay Gidey Hailu (2015), compare different data mining techniques to predict HIV/ AIDS test by using a CRISP-DM methodology and finds the best algorithm that used to effectively predict HIV/AIDS test. He tested Four popular data mining algorithms (Decision tree, Naive Bayes, Neural network, logistic regression) were used to build the model that predicts whether an individual was being tested for HIV. The final experimentation results indicated that the decision tree (random tree algorithm) performed the best with accuracy of 96%, the decision tree induction method (J48) came out to be the second best with a classification accuracy of 79%, followed by neural network (78%). Logistic regression has also achieved the least classification accuracy of 74%.

Another research was conducted on a title called "Prediction percentage of severity in HIV Patients using data mining algorithm" by Joglekar et al (), proposed a new algorithm by applying the same methodology with tesfaye Gidey for identifying HIV infections in patients and helpful in identifying the presence of HIV infection in a patient. As a result, medical conclusions, treatment procedures and decisions can be made by practitioners accurately.

However, the problem is all those previous studies were conducted by using a CRISP-DM methodology with small proportion of the dataset and measure the performance of the model only based on accuracy and also they used only percentage split method to classify the dataset into training and testing dataset, so this doesn't totally solve the problem that faced during the prognosis of HIV/AIDS. In this research the researcher would use a hybrid methodology and compare the model performances based on accuracy, sensitivity, and specificity to select the best

algorithm that predict HIV/AIDS based and also the researcher would used percentage split and K-fold cross validation methods to divide the dataset into training and test datasets to solve the above problems and attempt to answer the following research questions:

- Which algorithm is effective to prognosis HIV/AIDS?
- Which data mining algorithm is more effective to identify or predict the future status of the patient?

1.3. Objective of the study

1.3.1. General objective

The main objective of this study is to design/develop an effective HIV/AIDS prognosis model by using data mining techniques.

1.3.2. Specific objectives

- To prepare the data, for model building, by extracting and transforming the data into a format required for the data mining algorithm.
- To develop data mining model for HIV/AIDS prognosis
- To compare the performance of the model with other models.
- To compare the accuracy of the models
- To compare the sensitivity of the models
- To compare the specificity of the models
- To identify the drawbacks of the new models
- Report the result and forward recommendations.

1.4. Scope and limitation

The study is intended to design a model for effective HIV/AIDS prognosis.

The main limitation of this study was getting the data. Getting health related data for data mining researches was a very difficult problem even though the ethical considerations are made.

1.5. Significance of the Study

The Research would have the following Significance or benefits to the health care professionals to diagnosis the disease:

- It helps the professionals to diagnosis, pass effective decision and treat the disease
- It advice the health care workers to have know how about the diagnosis and treatment of HIV/AIDS.
- It helps the health care stake holders (planners, policy makers, and decision makers) to support their decision to plan and implement effective health policy and improve patients' diagnosis in the Healthcares.
- The study would serves as the input for future research works related with prognosis.

1.6. Organization of the research

This study is organized into six chapters. The first chapter briefly describes background to the problem area, and states the problem, objective of the study, scope and significance of the output of the research.

The second chapter deals with literature review about DM technology, tasks of DM, and its application in health care industry.

The third chapter provides discussions about the data mining methodologies, DM methods and techniques and algorithms.

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

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

In the sixth chapter conclusions and recommendations were presented

Chapter two

Literature review

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

2.1. Over views of data mining and knowledge discovery

Data mining derives its name from the similarities between searching for valuable business information in a large database, and mining a mountain for a vein of valuable ore. Data mining can generate new business opportunities by providing automated prediction of trends and behaviors, and discovery of previously unknown patterns.

Finding useful patterns in data has been given a variety of names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Many people treat data mining as a synonym for another popularly used term Knowledge discovery in database (KDD) and Data Mining is also just one step of the Knowledge Discovery Process (Cios et al, 2007).

One of the strengths of data mining, as opposed to more traditional statistical methods, is that it is not necessarily to know exactly what we are looking for before we start. Data mining uses powerful analytic tools to quickly and thoroughly explore mountains of data and pull out valuable and usable information. The primary use of data mining is to find something new in the data to discover a new piece of information that no one knew previously. This is data driven because it performs its processes on the data and then build theories and models based on discovered patterns or trends.

DM is not about analyzing small data sets that can be easily dealt with using many standard techniques rather it uses massive amount of data stored in files, data bases, and other repositories, and increasingly important to develop powerful tool for analysis and interpretation of the data and extracting the interesting hidden knowledge that could be useful for making appropriate decision (Cios et al, 2007).

The developments in digital data acquisition and storage technology have resulted in the growth of huge databases. The large size and complexity of such data in many scientific domains becomes different to manually analyze, explore, and understand data.

To undertake large data analysis projects, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems. In the current technology era lack of data is not the main problem rather generating useful knowledge is from huge data is the problem. Thus there is an urgent need for a new generation of computational theories and tools

to assist humans in extracting useful knowledge from the rapidly growing volumes of digital data and leads to the emerging of a new field called knowledge discovery in database and data mining (Fayyad et al., 1996).

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful means for analysis and interpretation of such data for the extraction of knowledge that which is important to support decision making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process (Zaïane, 1999).

The basic problem addressed by the KDD process is the mapping of low level data, which are too voluminous to understand and digest easily, into other forms that might be more compact, more abstract, or more useful. At the core of the process is the application of specific data-mining methods for pattern discovery and extraction (Fayyad, 1996), cited in (Denekew, 2003).

Data Mining is the process of extraction of useful knowledge from huge amounts of data to predict using techniques such as classification, clustering and association. A data mining system may generate numerous patterns. The discovered patterns can be similar to prior knowledge or expectations. Data Mining has two primary goals: Prediction and Description. Prediction involves variables in the dataset to predict unknown values or future values whereas Description focuses on finding patterns that describes the data interpreted by humans (Thiruma and Nagarajan, 2015).

Data mining refers to the analysis of the large quantities of data that are stored in computers and requires identification of a problem, along with collection of data that can lead to better understanding and computer models to provide statistical or other means of analysis (Olson and Delen, 2008). Data mining involves statistical and/or artificial intelligence analysis, usually applied to large-scale data sets.

2.2. Data mining processes

A typical data mining process includes data acquisition, data pre-processing, model building and model validation (Deshpande & Thakare, 2010).

Data acquisition

The first step in data mining is to select the types of data to be used. Although a target data set has been created for discovery in some applications, DM can be performed on a set of variables or data samples in a larger database called training set to create and model while holding back some of the data sets which are called test dataset for latter validation of the model.

Data preprocessing

Once the target data is selected, the data is then pre-processed for cleaning and transforming to improve the effectiveness of discovery. During this step, researchers remove the noise or outlier if necessary and decide on strategies for dealing with missing data fields. Then data is transformed to reduce the number of variables by converting one type of data to another such as numeric ones into categorical or deriving new attributes.

Model Building

The third step of data mining refers to a series of activities such as deciding on the type of data mining operations, selecting the data mining algorithms and mining the data. First, the type of the data mining operation such as, classification, regression, clustering, association rule discovery, and segmentation and deviation detection must be chosen. Based on the operations chosen for the application, an appropriate data mining technique is then selected based on the nature of the knowledge to be mined. The next step is selecting a particular algorithm within the data mining technique chosen. Choosing a data mining algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining technique with the overall objective of data mining. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

Interpretation and model evaluation

The fourth step of data mining process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understanding terms by users. In the interpretation of results, the researcher determines and resolves potential conflicts with previously known or decides redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal.

2.3. Data mining models

There are various data mining tasks that can be applied to solve a problem.
Depending on the nature of the problem, one or more different tasks can be combined to solve it because all work in a different way. Generally Data mining tasks are mainly classified as Predictive and Descriptive models depending on the use of data mining result (Durairaj and Ranjani, 2013)

Predictive modeling

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, regression, prediction and time series analysis are some examples of predictive modeling. As Durairaj and Ranjani (2013) indicated many of the data mining

applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefine groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable.

Classification modeling

In supervised learning, classification refers to the mapping of data items into one of the pre - defined classes. In the development of data mining tools that use statistical approaches, one of the critical tasks is to create a classification model, known as a classifier, which will predict the class of some entities or patterns based on the values of the input attributes. Choosing the right classifier is a critical step in the pattern recognition process. A variety of techniques have been used to obtain good classifiers. Some of the more widely used and well known classification algorithms are Naive Bayes, Decision Trees, Neural Network and Logistic Regression (Durairaj and Ranjani, 2013).

Naive Bayes algorithm

Naive Bayes is the simplest of the classification algorithms. It builds patterns by counting the correlations between all different states of the input attributes and all different states of the output attributes. Attributes can only have discrete values. Naive Bayes is based on Bayes' theorems and is naive in the sense that it does not take possible dependencies among the input attributes into account. Strong dependencies among the input attribute can therefore bias the identified patterns. Naïve Bayes is often used in the beginning of the data mining process to quickly explore the data but can also be a powerful predictor in some situations.

Decision Trees algorithm

Decision Trees can handle both discrete and continuous attributes but bins the continuous values if appropriate. The algorithm works recursively to build a tree that afterwards can be used for prediction. It searches for the input attribute that most cleanly divides the data across the states of the output attribute. That input attribute is used to split the data into subsets and then the same procedure is repeated for each of the subsets and so forth. When a new case of data is to be classified it is compared to the splits of the built tree thus creating a path from the root to a leaf node. That leaf node contains the predicted state of the output attribute. During training the tree is pruned using two algorithm parameters so that the resulting tree is not too deep which may cause over training. Very deep trees tend to over represent the training data instead of generalizing rules from it, which may result in a bad performance when classifying new cases of data. Decision Trees is one of the most popular algorithms because it is fast, easily understood and accurate if used properly.

Neural Network algorithm

The Neural Network algorithm is an artificial neural network that mimics the way the human mind works when presented with a problem. It analyzes all possible combinations of inputs and outputs and assigns weights to their relationships. It also looks for combinations of inputs that correlate to an output even though the inputs alone do not. There is also a hidden layer of nodes between the inputs and outputs, so that the inputs do not have to be directly correlated to an output. Instead the inputs can be related to a node in the hidden layer which in turn is related to an output. The resulting network can be used for prediction of new cases of data. Neural Network is suitable when trying to detect very complex relationships between inputs and outputs. However, the patterns extracted from the model are not well suited for exploration because they are hard to interpret.

Logistic Regression algorithm

Logistic Regression is a special case of the Neural Network algorithm in the way that it contains no hidden layer,, but besides that they are identical and therefore behave similarly. The removed hidden layer does not necessarily make it a weaker algorithm when it comes to predicting new cases of data. In some situations it can even perform better than Neural Network because the reduced complexity implies less risk of overtraining. Both Neural Network and Logistic Regression are able to handle both discrete and continuous attributes.

Knowledge discovery in database (KDD) process

The first KDD process was proposed by Fayyad in 1996 (Azevedo, and Filipe, 2008). This process consists of several steps that can be executed iteratively. KDD has been more formally defined as it is non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is the process of knowledge discovery while data mining is a technique applied for knowledge discovery considered as a step from the entire process (Azevedo, and Filipe, 2008). Generally KDD has five steps.

1. **Data Selection:** this stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
2. **Data Pre-processing**: this stage consists on the target data cleaning and pre-processing in order to obtain consistent data.
3. **Data Transformation**: this stage consists on the transformation of the data using dimensionality reduction or transformation methods.
4. **Data Mining**: this stage consists on the searching for patterns of interest in a particular representational form, depending on the DM objective mostly prediction.
5. **Interpretation/Evaluation**: this stage consists on the interpretation and evaluation of the mined patterns.

2.4. Application of data mining

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

Companies in the financial sector use data mining to determine market and industry characteristics as well as to predict individual company and stock performance, Predict which customers buy new policies ;Identify behavior patterns of risky customers (Olson and Delen, 2008).

There are a number of researches done to apply data mining techniques in health care domain in general to diagnosis and prognosis HIV/AIDS to extract hidden patterns and rule to support the decision makers.

2.4.1. Application of data mining in Health care

Data mining has great importance for area of medicine, and it represents comprehensive process that demands thorough understanding of needs of the Healthcare organizations. Knowledge gained with the use of techniques of data mining can be used to make successful decisions that will improve success of Healthcare organization and health of the patients. Data mining requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results. Data mining, once started, represents continuous cycle of knowledge discovery. For organizations, it presents one of the key things that help create a good business strategy (Asha et al, 2013).

Today, there have been many efforts with the goal of successful application of data mining in the Healthcare institutions. Primary potential of this technique lies in the possibility for research of hidden patterns in data sets in Healthcare domain. These patterns can be used for clinical diagnosis. However, available raw medical data are widely distributed, different and voluminous by nature. These data must be collected and stored in data warehouses in organized forms, and they can be integrated in order to form hospital information system.

Data mining technology provides customer oriented approach towards new and hidden patterns in data, from which the knowledge is being generated, the knowledge that can help in providing of medical and other services to the patients. Healthcare institutions that use data mining applications have the possibility to predict future requests, needs, desires, and conditions of the patients and to make adequate and optimal decisions about their treatments. With the future development of information communication technologies, data mining will achieve its full potential in the discovery of knowledge hidden in the medical data.

Data mining techniques can be implemented in hardware and software to add values for existing information and can be integrated with new products and systems. Disease prediction plays an important application in data mining.

Healthcare is an area with great potential for successful data mining because of the wealth of data available. However there is generally still a lack of effective analysis tools to maximize the utility of the data the Healthcare environment is often said to be information rich but knowledge poor

Health care industry generates large amounts of complex data's such as patient history, hospital resources, electronic records, information about medical devices etc. These data's serves as a key resource to process and analyze for knowledge extraction that enables the decision making and to save cost (Pradhan, 2014).

In general, the objectives of Data Miming usage in medicine can be generalized into two main groups: treatment resources optimization (Healthcare management domain), and treatment quality improvement (medical treatment and research domains) (niakšu, 2015).

Researches using data mining techniques have been applied in prognosis of various diseases such as cardiovascular diseases, AIDS, TB, diabetes and asthma.

Various data mining techniques such as Naïve Bayes classifier, Decision Tree, Support Vector Machines, and k Nearest Neighbor have been used to prognosis different diseases.

Now a day, applying data mining techniques to prognosis diseases in medical field becomes a typical task. Many researchers have been made in the past decade to diagnose diseases.

Mariammal, Jayanthi and Patra (2014), Proposed model that Diagnosis and Treatment Suggestion System Using Data Mining Techniques. They try to develop a model that diagnosis and treat all diseases by applying single and multiple data mining techniques.

Another study was conducted by Thirumal and Nagarajan (2015), propose a model to diagnosis of diabetes mellitus by Appling data mining techniques. The study was conducted on Pima Indians diabetes data set database and applies classification by using Naive baayes, C4.5 and support vector machine algorithms to predict the disease.

The other study was conducted by Rosma et al. (2013) they proposed a model that predicts AIDS Survival by using data mining approach. The study was used an adaptive fuzzy regression technique, FuReA, to predict the length of survival of AIDS patients based on their CD4, CD8 and viral load counts and then the predictive ability of the technique was compared with fuzzy neural network prediction models, and found that both FuReA and fuzzy neural network models were able to predict the survival of AIDS with an accuracy of 60% to 100% based on selected dependent variables.

Lakshmi et al. (2014) suggested a new model that Predicts the percentage of severity in HIV Patients using data mining algorithm. The study proposed an innovative data mining algorithm which helps doctors or trainer doctors to identify the severity level of HIV/AIDS in patients.

Moreover, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, and fraud detection, medical and scientific discovery.

When the data sets have been grown in size and complexity, direct hands on data analysis has increasingly been augmented with indirect, automatic data processing. This has been supported by applying data mining algorithms, such as neural networks, clustering, genetic algorithms, decision trees and support vector machines. Data mining is the process of applying these methods to data with the intention of extracting the hidden knowledge. A common task in medicine is thus classification using predictive models. Therefore applying data mining techniques on prognosis HIV/AIDS may provide a good result.

Chapter three

Methodology

3.1. Research Methodology

Methodology is the steps or procedure that the researcher follows to achieve the stated objectives. It is a road map that shows the direction how the researcher is going to conduct the research to reach the end.

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

3.2. Data mining methodologies

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

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

Many data mining process methodologies are available. However, the various steps do not differ much from methodology to methodology. According to Santos & Azevedo (2008) the followings are the most popular methodologies used by data mining tools; Knowledge Discovery in Database (KDD), Cross Industry Standard Process for Data Mining(CRISP-DM), Sample, Explore, Modify, Model, Assess(SEMMA) and Hybrid methodology. Brief description of each method is given as follows:

3.2.1. Knowledge discovery in database (KDD) process

According to Azevedo & Santos (2008) KDD is the traditional process of using DM methods to extract the hidden knowledge based on the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. Generally KDD process has five steps:

1. **Data Selection**: this step consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

2. **Data Pre-processing**: this step consists on t he target data cleaning and pre-processing in order to obtain consistent data.
3. **Data Transformation:** this step consists on the transformation of the data using dimensionality reduction or transformation methods.
4. **Data Mining:** this step consists on the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction).
5. **Interpretation/Evaluation**: this step consists on the interpretation and evaluation of the mined patterns.

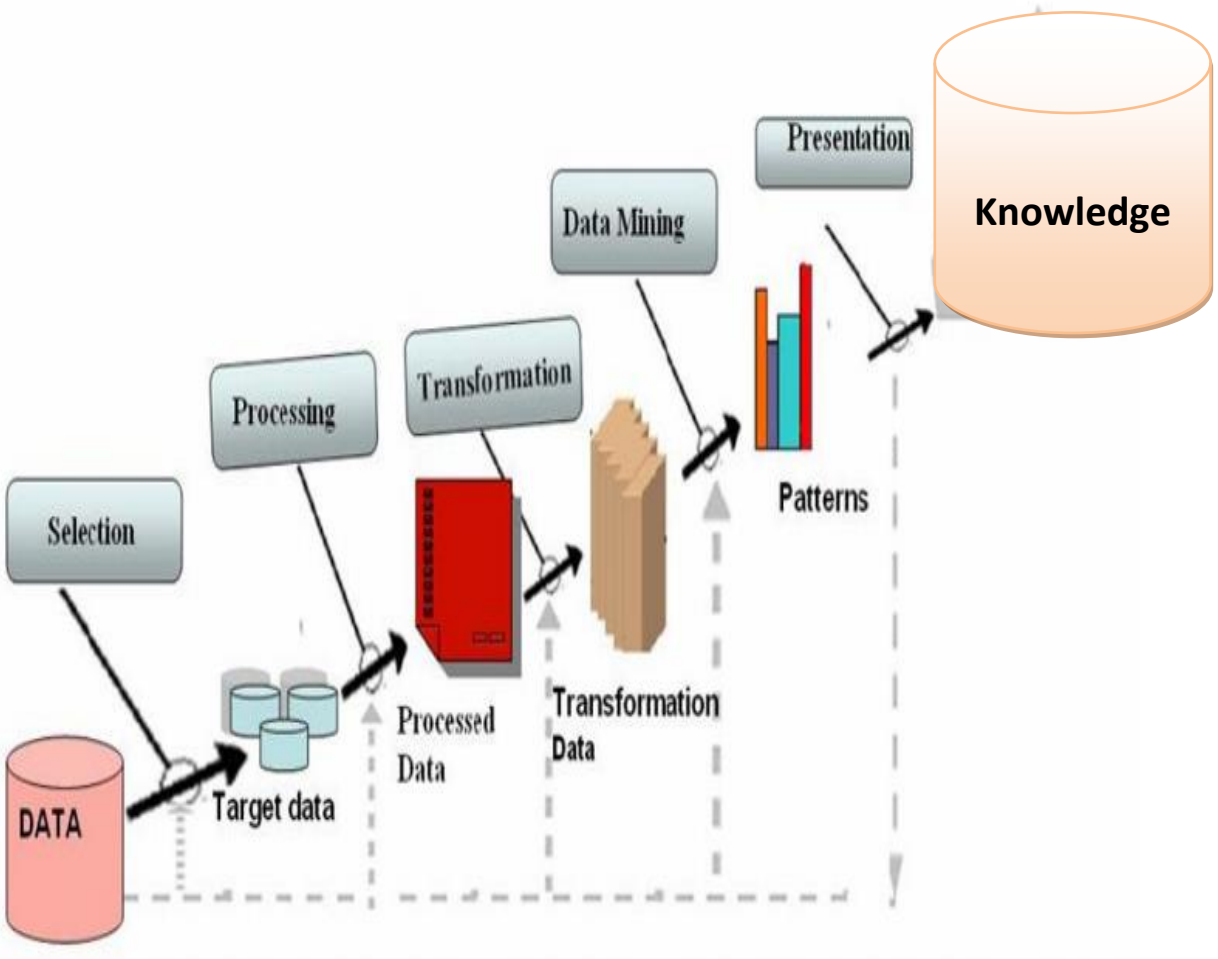

Figure 1 the KDD process (source: Azevado A. And Santos F., 2008, 'KDD, SEMMA AND CRISP-DM: A Parallel Overview)

The KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It also must be continued by the knowledge consolidation by incorporating this knowledge into the system (Fayyad et al, 1996).

3.2.2. The CRISP DM process

The CRISP-DM process was developed by the effort of a consortium initially composed with Daimler Chryrler, SPSS and NCR. CRISP-DM stands for Cross Industry Standard Process for Data Mining (Cios et al, 2007).

CRISP-DM (Cross Industry Standard Process for Data Mining) is a data mining project compromises a multi-step, iterative process that consist six steps/stages.

1. **Business understanding**

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2. **Data understanding**

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. **Data preparation**

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

4. **Modeling**

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

5. **Evaluation**

At this stage in the project you have built a model (or models) that appear to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.

A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

6. **Deployment**

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and

presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision making processes, for example in real time personalization of Web pages or repeated scoring of marketing databases. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created model.

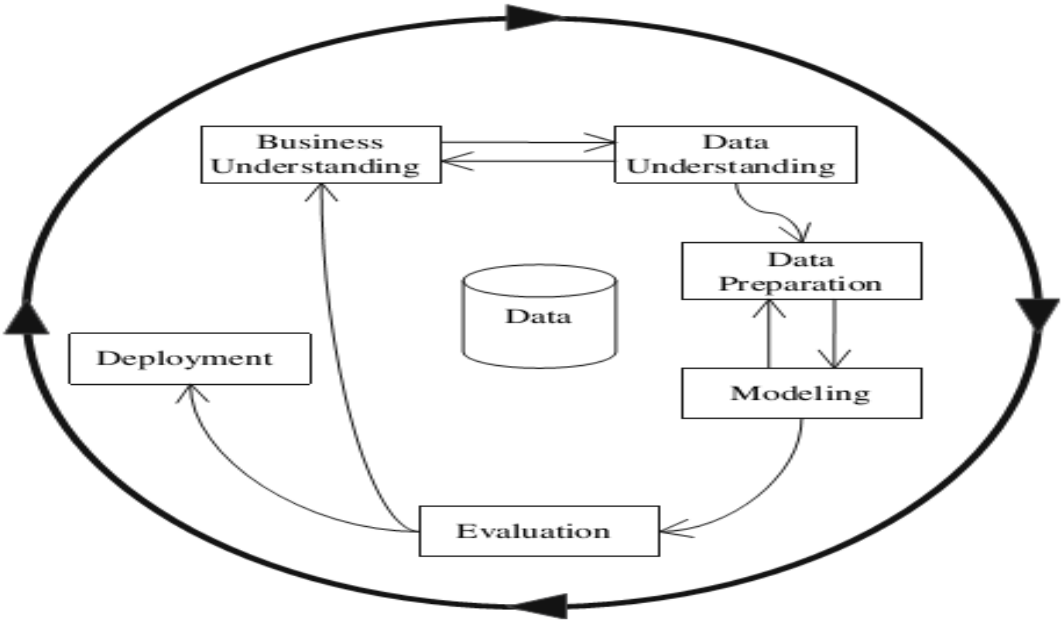

Figure 2The CRISP-DM KD process model (source: http://www.crisp-dm.org)

The CRISP-DM model is characterized by an easy-to-understand vocabulary and good documentation. It divides all steps into sub steps that provide all necessary details. It also acknowledges the strong iterative nature of the process, with loops between several of the steps. In general, it is a very successful and extensively applied model, mainly due to its grounding in practical, industrial, real-world knowledge discovery experience. The CRISP-DM model has been used in domains such as medicine, engineering, marketing, and sales (Cios et al., 2007).

3.2.3. The SEMMA Process

The SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. The SAS Institute considers a cycle with 5 stages for the process. By assessing the results gained from each stage of the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data.

1. **Sample**: This stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.
2. **Explore:** This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
3. **Modify**: This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
4. **Model:** This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
5. **Assess:** This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

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

3.2.4. Hybrid Models

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

1. **Understanding of the problem domain**. This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

2. **Understanding of the data**. This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing

values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

3. **Preparation of the data.** This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. discrete The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (by discreteization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.

4. **Data mining.** Here the data miner uses various DM methods to derive knowledge from preprocessed data.

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

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

Cios et al, 2007 described that hybrid model has lots of advantages when compared it with other methodologies. The main difference and extension of the hybrid methodology is:

- It providing more general, research-oriented description of the steps,
- It introducing a data mining step instead of the modeling step,
- It introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
- It modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

The hybrid model has been mostly used in medicine and software development areas.

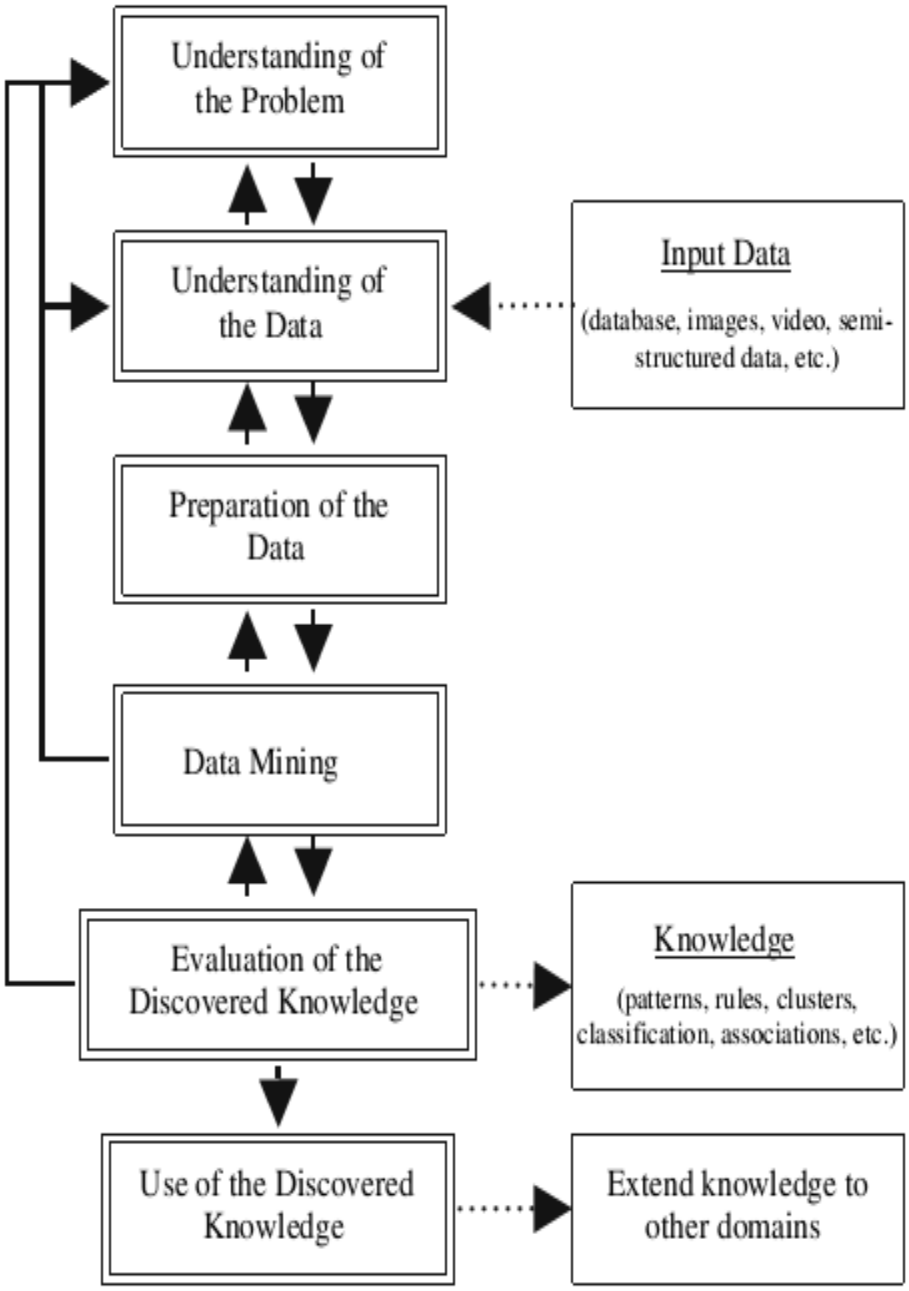

Figure 3 hybrid model (source: Cios et al, 2007, A Knowledge Discovery Approach)

3.3. Comparison of SEMMA, KDD and CRISP-DM processes

Today, research efforts have been focused on proposing new models, rather than improving design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made. The subsequent

models provide more generic and appropriate descriptions. Most of them are not tied specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application (kurgan and Musilek, 2006). Santos & Azevedo (2008) have summarized the association of the three most popular process model steps as shown in Table 1.

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

Table 1 Comparison of KDD, SEMMA and CRISP-DM Data mining Methodologies

The above table (table 1) illustrates that most of the steps to be followed in all methods look like similar. However, KDD and SEMMA do not have a step understanding the problem before selection and sample steps respectively. Even if, understanding the domain problem is often a prerequisite step to accomplish the project successfully the above three methodologies doesn't contain this step. So based on the above comparison we can conclude that SEMMA and CRISP-DM can be considered as an implementation of the KDD process developed by Fayyad et al in 1996 (Cios et al, 2007).

We emphasize that there is no universally "best" data mining methodology. Each of the methodologies has its strong and weak sides based on the application domain and particular objectives of the project/research (Cios et al, 2007).

In general this research can be conducted based on understanding of the domain problem and business context, consequently, the best fit data mining methodology to conduct this research is necessarily hybrid methodology.

This methodology has been chosen because the first reason is since it is a hybrid of both for academic and industrial purpose so since this research was conducted for academic purpose that is why hybrid methodology is selected the second reason is since it has more feedback stages/steps than the other methodologies. According to Cios et al (2007), the hybrid model has lots of advantages when compared it with other methodologies. The main differences and extensions include:

- providing more general, research-oriented description of the steps,

- Introducing a data mining step instead of the modeling step,
- Introducing several new explicit feedback mechanisms, (the CRISP -DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
- Modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

3.4. Data Mining Methods and techniques

Depending on their purpose of application, DM methods are divided into two groups: methods for prediction and methods for data characterization. Characterization tasks are aimed at finding patterns and associations, while prediction tasks are meant to predict certain events or certain unknown values within the relevant sphere of interests. The main methodological difference is that prediction requires a specific variable (class) to be included into the primary data. The solution can be numeric or categorical; respectively, DM methods for prediction are divided into regression and classification (Niaksu, 2015).

3.4.1. Predictive model

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, regression, prediction and time series analysis are some examples of predictive modeling. As Durairaj and Ranjani (2013) indicated many of the data mining applications are aimed to predict the future state of the data.

Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state.

Classification is a technique of mapping the target data to the predefine groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable.

3.4.2. Classification model

In supervised learning, classification refers to the mapping of data items into one of the pre - defined classes. In the development of data mining tools that use statistical approaches, one of the critical tasks is to create a classification model, known as a classifier, which will predict the class of some entities or patterns based on the values of the input attributes. Choosing the right classifier is a critical step in the pattern recognition process. A variety of techniques have been used to obtain good classifiers. Some of the more widely used and well known classification algorithms are Naive Bayes, Decision Trees, Neural Network and Logistic Regression (Durairaj and Ranjani, 2013)

Chapter four

Data Understanding and Preprocessing

4.1. Introduction

Data pre-processing is an often neglected but important step in the data mining process. The phrase "Garbage in Garbage out" is particularly applicable to data mining and machine learning. Analyzing the data that has not been carefully screened for data mining problems can produced misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection and so on. The product of data pre-processing is the final training set.

4.2. Overview of Ethiopian demographic and health survey (EDHS)

Demographic and Health Survey (DHS) in Ethiopia have been collected every five years since 2000. It is collected three times, 2000, 2005 and 2011.

The 2011 Ethiopia Demographic and Health Survey (2011 EDHS) is part of the worldwide MEASURE DHS project which is funded by the United States Agency for International Development (USAID) (EDHS,2011). The survey was implemented by the Ethiopian Central Statistical Agency (CSA). The funding for the EDHS was provided by the HIV/AIDS Prevention and Control Office (HAPCO), USAID, the United Nations Population Fund (UNFPA), the United Kingdom for International Development (DFID), the United Nations Children's Fund (UNICEF) and the Centers for Disease Control and Prevention (CDC). ICF International provided technical assistance through the MEASURE DHS project. The opinions expressed herein are those of the authors and do not necessarily reflect the views of USAID (F. Provost, 2010).

The 2011 Ethiopia Demographic and Health Survey (EDHS) is a nationally representative survey of 14,070 women age 15-49 and 6,033 men age 15-59. The EDHS is the second comprehensive survey conducted in Ethiopia as part of the worldwide Demographic and Health Surveys (DHS) project. The primary purpose of the EDHS is to furnish policymakers and planners with detailed information on fertility, family planning, infant, child, adult and maternal mortality, maternal and child health, nutrition and knowledge of HIV/AIDS and other sexually transmitted infections. In addition, in one of two households selected for the survey, women age 15-49 and children age 6-59 months were tested for anemia, and women age 15-49 and men age 15-59 were tested for HIV. The 2011 EDHS is the third survey in Ethiopia to provide population-based prevalence estimates for anemia and HIV (CSA, 2011).

4.3. Understanding of the problem domain

Understanding problem domain is working closely with domain experts to define the problem and determine the study goals, identifying key people, and learning about current solutions to the problem (Cios et.al, 2007).

The problem domain of this study was clearly defined in chapter1on the statement of the problem section. After understanding the problem to be addressed, the next step was analyzing and understanding the available data. The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of the available data (Cios et al, 2007).

4.4. Understanding of the data:

According to Cios et al. (2007) model, the next phase after understanding the business and problem domain has been data understanding. Hence, a prerequisite for undertaking DM research was dealing with data itself.

The Ethiopian 2011's EDHS HIV test raw datasets contains 14786 instances with 22 attributes. It is collected from CSA which is saved on SPSS as .sav format; and the researcher have converted into .xls format to make it readable on spread sheet application software and finally, I have converted it into .arff, format in order to make it ready to readable on WEKA software for analysis purpose.

The attributes of the raw data are:- 15 characters cluster/hh/line from recode, Cluster number, Household number, Line number of respondent, Sex, Age, Age 15-17, 18+, Line number of parent/responsible, Consent statement to parent, Consent to respondent, Sample result, Slept last night, Weight for HIV sample, Samples / test In lab file, Final result of all testing, Region, Type of place of residence, Result of individual interview, Level of education, Grade completed at HIVEDUC, Imputed age, and Sequence order in questionnaire.

4.4.1. Data Source

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

This study was based on data from the 2011 EDHS; the most recent national dataset on HIV testing that is available (as of January 2012). The 2011 EDHS included a nationally representative sample of women (aged 15- 49 years old) and men (aged 15 - 59 years old) from all eleven administrative regions in the country. The 2011 Ethiopia Demographic and Health Survey (EDHS) were conducted by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health. The Ethiopian Health and Nutrition Research Institute (EHNRI) were responsible for the testing of HIV from the dried blood samples (DBS). This is the third Demographic and Health Survey (DHS) conducted in Ethiopia, under the worldwide MEASURE DHS project, a USAID funded project providing support and technical assistance in the

implementation of population and health surveys in countries worldwide. The three EDHS surveys have been conducted at five-year intervals since 2000, and the 2011 EDHS is the second survey presenting results on HIV and anemia prevalence.

The 2011 Ethiopia DHS survey collected information on the population and health situation, covering topics on family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women's empowerment, and knowledge of HIV/AIDS were provided for the nine regional states and two city administrations. In addition, this report also provides data by urban and rural residence at the country level. In line with this, this study used the 2011 EDHS as a source of data especially on HIV testing result (ever been tested for HIIV) to predict whether an individual was being tested for HIV among adults in Ethiopia using data mining technology. The summary of the data sources is illustrated below in Table 2.

Source of data	Data coverage year	Number of records	Number of attributes	Size of the data	Data type
Central statics Agency (CSA)	2011 G.C	14786		1.06 MB	Nominal

Table 2 Description of data source and Number of records

4.4.2. Attribute selection

Not all attributes are relevant So, for selecting a subset of attributes relevant for mining, among all original attributes, attribute selection is required. Many irrelevant attributes may be present in data to be mined. So they need to be removed. Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid over fitting and improve model performance and to provide faster and more cost effective models.

The researcher has tried to consult the experts who are working in Ambo University College of medicine as lecturer in Nursing, Public health officer and pharmacy departments and the researcher have manually selected ten (10) attributes that are best suit for this study. Thus, the researcher has reasonably removed the other 13 attributes which are not selected for this study.

Table 3 shows the final attributes used for model building and their descript ion.

SN	ATTRIBUTES	DESCRIPTION	DATA TYPE
1	ID	Person's Serial numbers	Numeric
2	Sex	Person's sex	Nominal
3	Age	Person's age	Scale
4	Test_In_labFile	Whether the person is tested or not	Nominal
5	Final_testResult	Person's test result	Nominal
6	Level_of_Education	Person's education level	Nominal
7	marital status	Person's marital status	Nominal
8	NO of sex partner	Number of sexual partners a person have	Nominal
9	ever had sex	Whether the person ever had done sexual intercourse	Nominal
10	using condom during sex	Whether the person used condom during sexual intercourse	Nominal

4.4.2.1. Attributes Rank with Information Gain

The effect of the attributes on the model performance was investigated. The full training set containing a total of 14786 instances and 9 attributes. The attributes are selected by using InfoGainAttributeEval.

Table 4 shows that the final selected attributes by InfoGainAttributeEval and their rank.

Rank	Attribute
0.7793996	Test in labfile
0.0163523	Level of education
0.0092086	Sex
0.0016341	Age
0.0003976	Marital status
0.0003976	Number of sex partner
0.0000988	Ever had sex
0.000064	Using condom during sex
0	ID

The following figure (figure 4.1)illustrates that the depict ranked order of attribute based on their relevance for the reason that such attributes are very important for later experimentations by excluding the least relevant attributes.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 Final_testResult):
        Information Gain Ranking Filter

Ranked attributes:
 0.7793996   4 Test_In_labFile
 0.0163523   6 Level_of_Education
 0.0092086   2 Sex
 0.0016341   3 Age
 0.0003976   7 martial status
 0.0003976   8 NO of sex partner
 0.0000988   9 ever had sex
 0.000064   10 using condom during sex
 0           1 ID

Selected attributes: 4,6,2,3,7,8,9,10,1 : 9
```

Figure 4 Attribute Ranking

4.4.3. Data cleaning

Data cleaning refers to the pre-processing of data in order to remove or reduce noise and the treatment of missing values. It is the process of ensuring that all values in a dataset are consistent and correctly recorded. To do so all the data which are available on the database was cleaned to the same format. As a result the data were prepared for data analysis**.**

In the source data here are some instances with noisy data were filled erroneously, which brought mismatch between the raw SPSS dataset variable view and data view, so the researcher used MS-Excel application to clean the data.

4.4.4. Missing Values

The treatment of missing values is an important task in KDD process. Especially, while the dataset contains a large amount of missing data, the treatment of missing data can improve the quality of KDD dramatically. Selection of missing values series mean method highly depends on given data set, structure of attributes and missing data mechanism. Unfortunately missing data mechanism is usually unknown (Kaiser, 2014).

There are several strategies that could be used to handle missing values. Instances with missing values could be removed, missing values can be replaced with a certain value not present data can be replaced with a value that is representative for the data set. However all strategies

have their own flaws and which one to choose has to be decided from case to case. A common method for continuous attributes is to replace the missing value with the mean value of instances with no missing values. In the same way nominal missing values can be replaced with the mode value.

4.4.5. Descriptive Statistical Summary of Selected attributes

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

Table 5 shows the Statistical summary of selected Attributes.

Attribute	Total number	Distinct values	Count	Missing values	
				Count	Percentage (%)
ID	14786	14786	14768	0	0
Sex	14786	Male	7346	0	0
		Female	7440		
Age	14786	More than 18	12791	12	0
		Less than 18	1995		
Test_In_labFile	14786	Yes	11381	12	0
		No	3404		
		Problem inlab	1		
Final_testResult	14786	Positive	219	20	0
		Negative	11163		
		Not tested	3404		
Level_of_Education	14786	No education	7454	9	0
		Primary	3867		
		Secondary	2929		
		Higher	536		
marital status	14786	Never married	811	3	0
		Married	10015		

		Divorced	2231		
		Widowed	1729		
NO of sex partner	14786	None	811	0	0
		One	10020		
		Two	1724		
		More than two	2231		
ever had sex	14786	Yes	13811	24	0
		No	975		
using condom during sex	14786	Yes	3960	0	0
		No	10826		

4.5. Handling outlier value

outliers are usually the unwanted entries which always affects the data in one or the other form and distorts the distribution of the data Vijendra &Shivani(2007). Sometimes it becomes necessary to keep even the outlier entries because it plays an important role in the data but in this study in order to achieve the main objective effectively the researcher delete the entire outlier a found in the data set.

4.5.1. Data integration

I have collected the raw data from single source which is collect and filed on SPSS.

4.5.2. Data reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same or almost the same analytical results. Elimination of all samples is possible only when large data sets are available, and missing values occur only in a small percentage of samples and when analysis of the complete examples will not lead to serious bias during the inference. Elimination of attributes with missing values during analysis is not possible solution if we are interested in making inferences about these attributes. Both approaches are wasteful procedures since they usually decrease the information content of the data (Kaiser, 2014). Thus, I have removed 63 instances. Additionally, these removed attributes were not required for the study. Therefore the total record selected for this study is 14786 with 10 attributes.

4.5.3. Data transformation

In data transformation, the data are transformed into the appropriate formats for the mining tool for mining. I have done the data transformation by using Microsoft excel and manually change the code given to the attribute into related and understandable name based on the entire source of data stored in SPSS.

It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data.

The collected data in .spss format is converted to Microsoft Excel 2007(.xls) format for performing some pre-processing techniques. This data format cannot directly processed by WEKA data mining tool. Thus, transformation is performed. The **.xls** format is saved in CSV (Comma Separated Value) see (Appendix 5).

4.6. Association Rules

Data mining functionalities were used to specify the kind of patterns to be found in data. Association Analysis is useful for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules or set of frequent items.

4.6.1. Apriori Approach

Association rules are one of the major data mining techniques. It is perhaps the most common form of local-pattern discovery in unsupervised learning systems (Setiono and Liu, 1996). Association rules are widely used in data mining to find patterns in data. Therefore, this study used association rules of data mining technique to examine which instances of HIV testing frequently occurred in a database and presents the patterns as rules among the records.

Association rule mining

```
Best rules found:

 1. Final_testResult=Negative 11163 ==> Test_In_labFile=Yes 11162    conf:(1)
 2. Final_testResult=Negative ever had sex=yes 10566 ==> Test_In_labFile=Yes 10565    conf:(1)
 3. Test_In_labFile=Yes ever had sex=yes 10770 ==> Final_testResult=Negative 10565    conf:(0.98)
 4. Test_In_labFile=Yes 11381 ==> Final_testResult=Negative 11162    conf:(0.98)
 5. Final_testResult=Negative 11163 ==> ever had sex=yes 10566    conf:(0.95)
 6. Test_In_labFile=Yes Final_testResult=Negative 11162 ==> ever had sex=yes 10565    conf:(0.95)
 7. Final_testResult=Negative 11163 ==> Test_In_labFile=Yes ever had sex=yes 10565    conf:(0.95)
 8. Test_In_labFile=Yes 11381 ==> ever had sex=yes 10770    conf:(0.95)
 9. Age=more than 18 12791 ==> ever had sex=yes 12083    conf:(0.94)
10. Test_In_labFile=Yes 11381 ==> Final_testResult=Negative ever had sex=yes 10565    conf:(0.93)
```

Figure 5 Association rule

Chapter five

Experimentation and analysis of results

5.1. Overview of Experimentation

In this chapter, the researcher describes the techniques that have been used to develop a model that prognosis HIV/AIDS. This research integrated the main stages that characterize a data mining process. This study has been organized according to hybrid processing model, which is used to develop the model. Here the researcher discuss the experimentation process by relating the steps followed ,the choice made , the task accomplished , the result obtained, evaluation of the model and results , and present it in a way that the organization can easily understand and use it.

Experiments have been carried to identify classification model and to extract relevant actionable association rules.

In this study different experiments were conducted using various data mining methods to derive knowledge from preprocessed data to prognosis HIV/AIDS from the preprocessed data. According to the methodology of this study after preparation of the data, the next task is the mining process. As it has been stated in the previous sections, a total of 14786 data were preprocessed to perform the experiment.

5.2. Data mining

5.2.1. Model Building

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

5.2.2. Selecting Modeling Technique

Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification techniques namely Bayes classification, decision tree classification and Artificial Neural network classification has been selected for model building.

The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, Naivebayes, J48 and neural network (Multilayer Perceptron) algorithms are used for experimentation of this study.

The reason why the researcher selected the above algorithms is, because the algorithms are easy to understand and interpret the results of the model and also have advantages such as high tolerance to noise, and the ability to classify unseen patterns.

5.3. Experimental Setup

In any data mining research and project before building a model, we should generate a procedure or mechanism to test the model performance. For instance, in the supervised data mining task such as classification, it is common to use classification accuracy measure, sensitivity, specificity, precision, recall, error rates and judgment of the experts are used as to measure the performance/test of the developed data mining model.

In this research 14,786 datasets are used for training and testing. WEKA 3.6.9 software has used to set up and measure the quality, validity and test of the selected model. The following figure (figure 5.1) shows that the WEKA GUI.

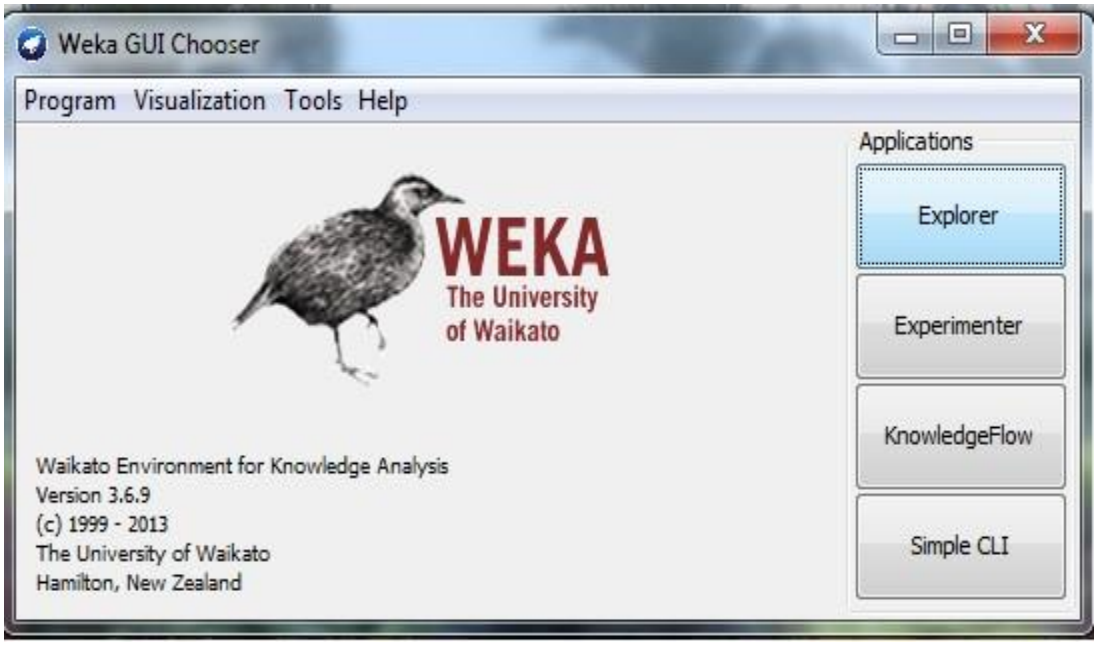

Figure 6 the WEKA GUI

In order to perform the experiment the researcher used two methods to classify the dataset as training and test datasets, the first method is k-fold (10-folds) cross validation and, the second method is percentage split.

Therefore the dataset are randomly partitioned equally into ten parts. Consequently, 90% of the dataset is used for training and 10 % for testing and the dataset are partitioned in to percentages splits option (70%:30%) that means 70% of the dataset is used for training and the remaining 30% for testing purpose.

5.4. Experimentation

AS I have described in section 5.2, model building is the crucial step In Hybrid methodology. Model building is not single process rather it is an iterative process's. Therefore in this study different experiments are conducted using three algorithms namely J48, Naivebayes and Artificial neural network algorithms for building the best predictive model.

As the researcher explained on the experiment setup section (section 5.3) to build the model for each algorithm I perform two experiments based on the two methods namely the experiment performed based on K-fold (10 – fold)cross validation method called experiment I and the second experiment performed based on the percentage split method called experiment II.

5.4.1. Model building using Naivebayes algorithm

- **Experiment I**

For this experiment I used the K-fold (10 – fold) cross validation test option to train and test the classification model. Table below (table 6) shows that the resulting confusion matrix of this model.

Actual	Predicted			Total	Correctly classified (Accuracy rate)	Sensitivity	Specificity
	Positive	Negative	Not_Tested				
Positive	0	219	0	219	0%		
Negative	0	11162	1	11163	99.9910%	**98.50%**	**98.1%**
Not_Tested	0	1	3403	3404	99.9706%		
Total	0	11382	3404	14786	**98.5053%**		

Table 6 Confusion Matrix output of the Naivebayes Algorithm with the K-fold (10-fold) Cross Validation method.

As shown on the confusion matrix table (table 6) the Naivebayes, learning algorithm scored an accuracy of 98.5053 % which indicates that out of 14786 total numbers of instances 14565 (98.5053%) instances are classified correctly and 221 (1.4947%) instances are incorrectly classified or misclassified,

Which means that out of the total 14786 instances 11163 instances are negative and 11162 (99.9910%) of them are classified correctly as negative and only 1 instance is incorrectly classified as not tested by the model and 3404 instances were indicated as Not tested and out of them 3403 (99.9706%) are correctly classified as not tested and the remaining 1 instance is misclassified as Negative by the model. At the same time, out of 219 positive instances none of them are correctly classified as positive, so long as all the positive instances are 100% incorrectly classified or misclassified as negative.

- **Experiment II**

For this experiment the K-fold cross validation method is changed into percentage split test option to train and test the classification model.

The Table below (table 7) shows that the resulting confusion matrix of the model developed by percentage split set to 70% by Naivebayes Algorithm.

Actual	Predicted			Total	Correctly classified (Accuracy rate)	Sensitivity	Specificity
	Positive	Negative	Not_Tested				
Positive	0	77	0	77	0%	**98.3%**	**97.6%**
Negative	0	3316	0	3316	100%		
Not_Tested	0	0	1043	1043	100%		
Total	0	3393	1043	4436	**98.2642%**		

Table 7 Confusion Matrix output of the Naivebayes Algorithm with the percentage split set70%.

Out of the 14786 total records 10350 (70%) of the instances were used as training dataset and the remaining, 4436 (30%) of the instances were used as a testing dataset. As shown in the above confusion matrix table (table 5.2), the Naivebayes learning algorithm scored an accuracy of 98.2642% that indicates out of 4436 total number of testing instances 4359 (98.2642 %) of them are classified correctly and the remaining 77 (1.7358%) testing instances are misclassified or incorrectly classified.

This means that out of 3316 Negative instances, all the 3316 (100%) of them are classified correctly as a Negative and out of 1043 Not tested records, 1043 (100%) of them are classified correctly as Not tested. At the same time, out of 77 positive instances none of them are correctly classified as positive, so long as all the positive instances are 100% incorrectly classified or misclassified as negative.

To conclude, the above two experiments namely experiment I and II performed in order to built the classification model using Naivebayes classification Algorithm by applying k-fold cross validation and percentage split method in respectively on the experiments.

Table 6 and table 7 shows that the prediction accuracy of the models based on the above two methods respectively. The first experiment was performed based on 10-fold cross validation method and predicts with 98.5053% accuracy rate, and the second experiment performed based on 70%:30% percentage split predicts with 98.2642% accuracy rate.

So to sum up, when we compared the two experiments the first experiment performed based on K-fold cross validation has a better accuracy performance than the second experiment performed by percentage split.

5.4.2. Model building using J48 Decision Tree algorithm

- **Experiment I**

For this experiment I used the K-fold (10 – fold) cross validation test option to train and test the classification model using J48 Decision tree algorithm. Table below (table 8) shows that the resulting confusion matrix of this model.

Actual	Predicted			Total	Correctly classified (Accuracy rate)	Sensitivity	Specificity
	Positive	Negative	Not_Tested				
Positive	0	219	0	219	0%	**98.51%**	**97.1%**
Negative	0	11163	0	11163	100%		
Not_Tested	0	1	3403	3404	99.9706%		
Total	0	11382	3403	14786	**98.5121%**		

Table 8 Confusion Matrix output of the J48 Algorithm with the K-fold (10-fold) Cross Validation method.

As shown on the confusion matrix table (table 8) the J48 learning algorithm scored an accuracy of 98.5121 % which indicates that out of 14786 total numbers of instances 14566 (98.5121%) instances are classified correctly and 220 (1.4879%) instances are incorrectly classified or misclassified,

Which means that out of the total 14786 instances 11163 instances are negative and 11163 (100%) of them are classified correctly as negative by the model and 3404 instances were

indicated as Not tested and out of them 3403 (99.9706%) are correctly classified as not tested and the remaining 1 instance is misclassified as Negative by the model. At the same time, out of 219 positive instances none of them are correctly classified as positive, so long as all the positive instances are 100% incorrectly classified or misclassified as negative.

- **Experiment II**

For this experiment the K-fold cross validation method is changed into percentage split test option to train and test the classification model. The Table below (table 9) shows that the resulting confusion matrix of the model developed by percentage split set to 70% by J48 Algorithm.

Actual	Predicted			Total	Correctly classified (Accuracy rate)	Sensitivity	Specificity
	Positive	Negative	Not_Tested				
Positive	0	77	0	77	0%	**98.3%**	**96.8%**
Negative	0	3316	0	3316	100%		
Not_Tested	0	0	1043	1043	100%		
Total	0	3393	1043	4436	**98.2642%**		

Table 9 Confusion Matrix output of the J48 Algorithm with the percentage split set70%.

Out of the 14786 total records 10350 (70%) of the instances were used as training dataset and the remaining, 4436 (30%) of the instances were used as a testing dataset. As shown in the above confusion matrix table (table 9), the J48 learning algorithm scored an accuracy of 98.2642% that indicates out of 4436 total number of testing instances 4359 (98.2642 %) of them are classified correctly and the remaining 77 (1.7358%) testing instances are misclassified or incorrectly classified.

This means that out of 3316 Negative instances, all the 3316 (100%) of them are classified correctly as a Negative and out of 1043 Not tested records, 1043 (100%) of them are classified correctly as Not tested. At the same time, out of 77 positive instances none of them are correctly classified as positive, so long as all the positive instances are 100% incorrectly classified or misclassified as negative.

To conclude, the above two experiments namely experiment I and II performed in order to built the classification model using J48 classification Algorithm by applying k-fold cross validation and percentage split methods in respectively on the experiments. The above two tables (table 8 and table 9) shows that the prediction accuracy of the models based on the two (k - fold cross validation and percentage split) methods respectively. The first experiment was performed based on 10-fold cross validation method and predicts with 98.5121% accuracy rate, and the second experiment performed based on 70%:30% percentage split predicts with 98.2642% accuracy rate.

So, when we compared the two experiments performed by different methods, the first experiment performed based on K-fold cross validation has a better accuracy performance than the second experiment performed by percentage split.

5.4.3. Model building using Neural Network Algorithm

- **Experiment I**

For this experiment I used the K-fold (10 – fold) cross validation test option to train and test the classification model using Neural Network (MLP) Algorithm.

Table below (table 10) shows that the resulting confusion matrix of this model.

Actual	Predicted			Total	Correctly classified (Accuracy rate)	Sensitivity	Specificity
	Positive	Negative	Not_Tested				
Positive	0	219	0	219	0%	**98.50%**	**98.1%**
Negative	0	11162	1	11163	99.9910%		
Not_Tested	0	1	3403	3404	99.9706%		
Total	0	11382	3404	14786	**98.5053%**		

Table 10 Confusion Matrix output of the Neural Network (MLP) Algorithm with the K-fold (10-fold) Cross Validation method.

As shown on the confusion matrix table (table 10) the Neural Network Multilayer perceptron Algorithm, learning algorithm scored an accuracy of 98.5053 % which indicates that out of 14786 total numbers of instances 14565 (98.5053%) instances are classified correctly and 221 (1.4947%) instances are incorrectly classified or misclassified,

Which means that out of the total 14786 instances 11163 instances are negative and 11162 (99.9910%) of them are classified correctly as negative and only 1 instance is incorrectly classified as not tested by the model and 3404 instances were indicated as Not tested and out of them 3403 (99.9706%) are correctly classified as not tested and the remaining 1 instance is misclassified as Negative by the model. At the same time, out of 219 positive instances none of them are correctly classified as positive, so long as all the positive instances are 100% incorrectly classified or misclassified as negative.

- **Experiment II**

For this experiment the K-fold cross validation method is changed into percentage split test option to train and test the classification model. The Table below (table 11) shows that the resulting confusion matrix of the model developed by percentage split set to 70% by Naivebayes Algorithm.

Actual	Predicted			Total	Correctly classified (Accuracy rate)	Sensitivity	Specificity
	Positive	Negative	Not_Tested				
Positive	0	77	0	77	0%	**98.3%**	**97.7%**
Negative	0	3316	0	3316	100%		
Not_Tested	0	0	1043	1043	100%		
Total	0	3393	1043	4436	**98.2642%**		

Table 11 Confusion Matrix output of Neural Network (MLP) Algorithm with the percentage split set70%.

Out of the 14786 total records 10350 (70%) of the instances were used as training dataset and the remaining, 4436 (30%) of the instances were used as a testing dataset. As shown in the above confusion matrix table (table 11), the Neural Network Multilayer perceptron, learning algorithm scored an accuracy of 98.2642% that indicates out of 4436 total number of testing instances 4359 (98.2642 %) of them are classified correctly and the remaining 77 (1.7358%) testing instances are misclassified or incorrectly classified.

This means that out of 3316 Negative instances, all the 3316 (100%) of them are classified correctly as a Negative and out of 1043 Not tested records, 1043 (100%) of them are classified correctly as Not tested. At the same time, out of 77 positive instances none of them are correctly

classified as positive, so long as all the positive instances are 100% incorrectly classified or misclassified as negative.

To conclude, the above two experiments namely experiment I and II performed in order to built the classification model using Neural Network Multilayer perceptron classification Algorithm by applying k-fold cross validation and percentage split method in respectively on the experiments. The above two tables (table 10 and table 11) showed that the prediction accuracy of the models based on the above two methods respectively. The first experiment was performed based on 10-fold cross validation method and predicts with 98.5053% accuracy rate, and the second experiment performed based on 70%:30% percentage split predicts with 98.2642% accuracy rate.

So to sum up, when we compared the two experiments the first experiment performed based on K-fold cross validation has a better accuracy performance than the second experiment performed by percentage split.

5.5. Performance Comparison of Naivebayes, J48 and Neural Network (MLP) models

Selecting a better classification technique for building a model, which performs best in handling the prediction, is one of the aims of this research. For this reason, the three selected classification model with respective best performance accuracy, sensitivity and specificity is listed in the below table (table 12).

Algorithms Used	***Time taken(sec.)***	***Accuracy (%)***	***Sensitivity (%)***	***Specificity (%)***
10 fold cross validation test option				
NaiveBayes	0.19	98.5053	98.50	98.1
J48	0.54	98.5121	98.51	97.1
MLP	160.38	98.5053	98.50	98.1
Percentage split test option				
NaiveBayes	0.17	98.2642	98.3	97.6

J48	0.48	98.2642	98.3	96.8
MLP	164.86	98.2642	98.3	97.7

Table 12 Comparison of the results of the models

The above experiments were performed on WEKA experimenter by using training sets in WEKA test option and the experiments were performed by using two methods namely 10-fold cross validation and percentage split test option. In these experiments three algorithms are used. The algorithms are Naïve Bayes, J48 decision tree and Neural Network (MLP). In order to develop the predictive model the three algorithms are tested based on the two methods, three models are developed in each methods and totally six models are developed based on the two methods.

As shown on the above comparison table (table 12), all the results were almost all closely equal but the difference lays on the execution period or the time taken to build the model.

5.6. Predictive Model Performance Evaluation Metrics

Evaluation of predictive data mining algorithms can be compared according to a number of measures. In comparing the performance of different predictive data mining algorithms to determine its predictability, some quantities that interpret the goodness of fit of a model, and error measurements must be considered. In clinical medicine, the ability to detect patients with sickness or exclude patients without sickness is often described by terms such as Accuracy, sensitivity, specificity, positive predictive value and negative predictive value. The True Positive Rate, False Positive Rate, True Negative Rate and False Negative Rate determine the predictive data mining efficiency. The sensitivity metrics is also called true positive rate or positive class accuracy, while specificity is referred to as true negative rate or negative class accuracy.

The learning algorithms were executed by 10 fold cross-validation and percentage split test options on six experiments. In this study, we used three predictive model performance evaluation metrics; accuracy, sensitivity, and Specificity.

The final comparative analysis of the models shows on the above table the J48 pruned tree classifier algorithm with 10-fold cross validation test option performed best classification accuracy of 98.5121% with 98.51% sensitivity, where as Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option performed best classification specificity of 98.1% and 98.1% result respectively.

So, to sum up the model which was developed by the J48 pruned tree classifier algorithm with 10-fold cross validation test option method was selected as the best prognosis model based on

the two (accuracy and sensitivity) evaluation methods used in this study. Based on specificity the third evaluation method used in this study the model that was developed by Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option method was selected as the best prognosis model.

When we compared Tesfaye's work with this research his final experimentation results indicated that the decision tree (random tree algorithm) performed the best with accuracy of 96%, the decision tree induction method (J48) came out to be the second best with a classification accuracy of 79%, followed by neural network (78%). Logistic regression has also achieved the least classification accuracy of 74%, but my final comparative analysis study indicates the model which was developed by the J48 pruned tree classifier algorithm with 10-fold cross validation test option method was selected as the best prognosis model based on the two (accuracy and sensitivity) evaluation methods used in this study. Based on specificity the third evaluation method used in this study the model that was developed by Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option method was selected as the best prognosis model.

5.7. Association Rules

Association rule mining was performed using apriori algorithm to discover the relationship of the selected attributes with being tested for HIV. Apriori in WEKA 3.6.9 starts with the upper bound support and incrementally decreases support by delta value. In most cases, it is sufficient to focus on a combination of support and confidence to quantitatively measure the quality of the rules. However, the real value of a rule, in terms of usefulness and action ability is subjective and depends heavily on the particular domain and business objectives. To conduct the association rules for this study, the machine used a minimum support of 70%% and with 90% of confidence level based on the attributes it took as inputs. The 10 best rules are annexed (Appendix 5).

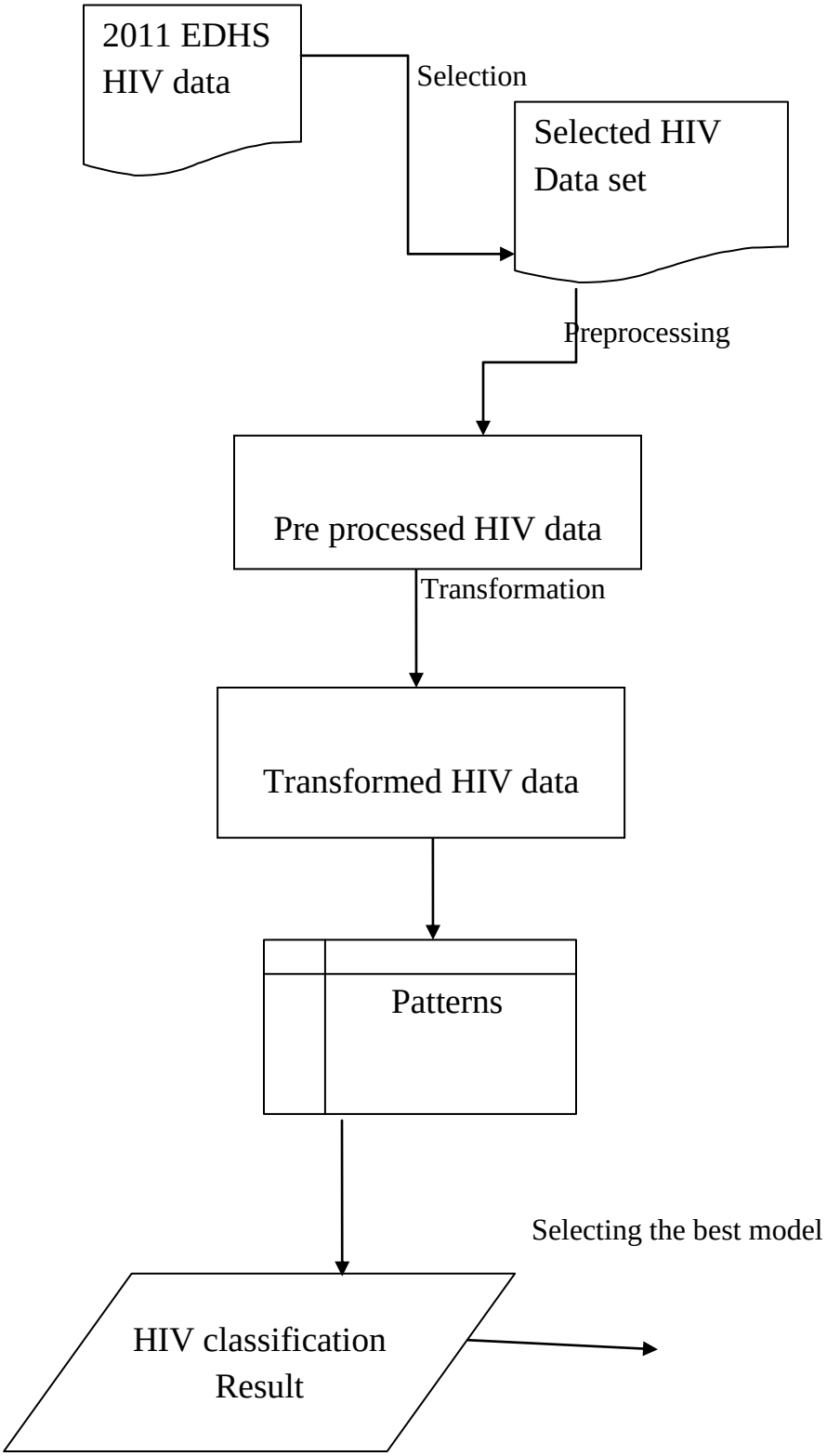

Figure 7 Architecture of HIV/AIDS testing classification data mining model

5.8. Using discovered knowledge

This is the final step of a hybrid methodology that shows the plan how and where to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed. Also this step is to put discovered knowledge in practical use either by documenting it and reporting it or by embedding it in a computer system. In first sight, this stage might be regarded as trivial and straightforward, but this is not the case. The conclusions drawn from the KDD process often reveal the complex nature of the problem and its solutions. This is not surprising as data mining techniques are not necessary when dealing with simple problem. Hence, the implementation of the new knowledge should often be done in gradually, while continuously monitoring the result achieved and the degree to which they fulfill the expectations.

The above generated results in this research are encouraging. This discovered knowledge could be used for classify HIV/AIDS testing results as positive, negative and not tested classes accordingly. In order to use these generated results effectively and efficiently the organization should have to first design a knowledge base system, which can provide advice for the domain experts to improve the decision making process.

CHAPTER SIX

Conclusion and recommendation

6.1. Conclusion

The effective use of information and technology is crucial for health care organizations to stay competitive in today's complex, evolving environment. The challenges faced when trying to make sense of large, diverse, and often complex data source are considerable. In an effort to turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficacy of patient care. Data mining can be used to help predict future patient behavior and to improve patients' treatment programs.

In this research, an attempt has been made to apply the data mining technology for effective prognosis of HIV/AIDS using EHDS HIV data set. Data mining technology is basically follows iterative process such as: Data collection, Data preparation and understanding, model building and evaluation. As I explained many different data mining methodologies in chapter two, there are possibilities of additional steps other than the above lists for instance KDD includes nine steps. The iterative nature of the process assist the data miner to back and forth at different and fix it where the problems are arise.

The main objective of this study is to design/develop an effective HIV/AIDS prognosis model by using data mining techniques. This investigation, conducted according to Cios et al (2007) hybrid data mining process model to achieve the main objective of this study. The Cios et al (2007) hybrid data mining model consists six steps namely understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge. The data used in this research has been gathered from Central Statistics Agency (CSA). After the data has been collected (14786 dataset), it has been preprocessed and prepared in a suitable format for the data mining tool to apply the data mining tasks on it.

The study was conducted using three classification techniques namely bayes, decision tree and Artificial Neural network. For model building and experimentation three algorithms are used namely NaiveBayes, J48 and MLP algorithms. In order to develop the predictive model the three algorithms were tested based on the 10-fold cross validation and percentage split test option methods, three models were developed in each methods, and totally six models were developed based on the two methods.

By changing the training test options and the default parameter values of the algorithm, these models are tested and evaluated. Finally, the J48 pruned tree classifier algorithm with 10-fold cross validation test option performed best classification accuracy of 98.5121% with 98.51% sensitivity, where as Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option performed best classification specificity of 98.1% and 98.1% result respectively.

Since, the model that developed by J48 decision tree algorithm with a 10-fold cross validation test option had scored better performance accuracy and sensitivity and Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option models had scored better performance specificity based on these evaluation parameters, it is the researcher's conclusion that J48 decision tree with 10-fold cross validation test option classification model has been an appropriate technique in terms of accuracy and sensitivity where as Naive Bayes and Neural Network (MLP) classifier algorithms with 10-fold cross validation test option models had been appropriate techniques in terms of specificity for this research on HIV/AIDS prognosis.

In general the results obtained from this research indicate that data mining is useful and appropriate tool in bringing relevant information to the service providers as well as decision and policy makers.

6.2. Recommendations

The researcher makes the following recommendations based on the findings of this study.

- This research has done for academic purpose and proved that J48 10-fold cross validation algorithm were discover hidden knowledge which was interesting and accepted by experts. But if knowledge based system were integrated on it, it may becomes a better advisory system for the health workers of the organization.
- The researcher used only two classification methods namely; 10-fold cross validation and percentage split classification methods. But, those data classification methods which were not tested on this research might result important patterns which were used to predict HIV/AIDS effectively and this might increase the performance of the model.
- To investigate this study, the researcher used imbalanced dataset to predict HIV/AIDS. However, researches can also be conducted using balanced data and other data mining techniques.

REFERENCESE

A Machine Learning Approach', (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 7, pp. 89 -94.

Asha T., S. Natarajan, K.N.B. Murthy (2011). 'A Data Mining Approach to the Diagnosis of Tuberculosis by cascading Clustering and Classification'.

Asha T., S. Natarajan, K.N.B. Murthy (2011). 'Effective Classification Algorithms to Predict the Accuracy of Tuberculosis:

Asia Nesredin (2012). Mining Patients' Data for Effective Tuberculosis Diagnosis: The Case of Menelik Ii Hospital. Master Thesis .Addis Ababa University School of Information Science.

Azevado A. and Santos F., 2008, 'KDD, SEMMA AND CRISP-DM: A Parallel Overview ', IADIS European Conference Data Mining, Portugal, pp. 182-185.

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

Bhasin, M. L. (2006). Data Mining: A Competitive Tool in the Banking and Retail Industries. *Banking and finance, 588*.

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

Denekew Abera Jembere (2003). The Application of Data Mining To Support Customer Relationship Management at Ethiopian Airlines.Master Thesis.Addis Ababa University School of Information Science.

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

Durairaj, M., & Ranjani, V. (2013). Data mining applications in Healthcare sector a study. *International Journal of Scientific and Technology Research*, *2*(10), 29-35.

Emir, S., Dincer, H., Hacioglu, U., & Yuksel, S. (2016). Comparative Study of Outlier Detection Algorithms via Fundamental Analysis Variables: An Application on Firms Listed in Borsa Istanbul. International Journal of Research in Business and Social Science (2147-4478), 4(4), 45-60.

Ethiopia Demographic and Health Survey (2005).

Ethiopia Demographic and Health Survey (2011).

F. Provost (2010). Machine Learning from Imbalanced Datasets.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

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

Hailu, T.G. (2015) Comparing Data Mining Techniques in HIV Testing Prediction. Intelligent Information Management, 7, 153-180. http://dx.doi.org/10.4236/iim.2015.7301

Huang, M. J., Chen, M. Y., & Lee, S. C. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*, *32*(3), 856-867.

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

Joglekar, A., Lakshmi, G. P., & Jani, M. Extraction of Rules For Predicting HIV Infections And Computing Support And Confidence.

Kozakura, S. Ogawa, H., Miura, H., Matsuda, N., Taki, H., Hori, S., & Abe, N. (2006, October). An interpretation method for classification trees in bio-data mining. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 620-627). Springer Berlin Heidelberg.

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

M. Durairaj and V. Ranjani (2013). Data Mining Applications in Healthcare Sector.

Madan L., 2006, 'Data Mining: A Competitive Tool in the Banking and Retail industries', the charter Accountant, pp. 588-594.

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

Mariammal.D, Jayanthi.S, Dr. P.S.K.Patra (2014). Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques.

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

Niaksu, O. (2015). CRISP Data Mining Methodology Extension for Medical Domain. Baltic Journal of Modern Computing, 3(2), 92.
Niakšu, O. Data Mining in Medicine: applications, challenges and possibilities. Redaktorių taryba, 19.
Obenshain, M. K. (2004). Application of data mining techniques to Healthcare data. *Infection Control & Hospital Epidemiology*, *25*(08), 690-695.
Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. Springer Science & Business Media.
Osmar R. Zaïane (1999). Principles of Knowledge Discovery in Databases.

Paliwal, P., & Malviya, M. An Efficient Method for Predicting Heart Disease Problem Using Fitness Value.

Saitta, S. (2008). *Data mining methodologies for supporting engineers during system identification* (Doctoral dissertation, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE).

Santos, Manuel Filipe, & Azevedo,Ana (2008).KDD, SEMMA AND CRISP-DM: a parallel overview .IADIS European Conference data miing.p.182-187.

Smolander, K., Tahvanainen, V.P., Lyytinen, K., (1990) How to Combine Tools and Methods in Practice a Field Study. In: Lecture Notes in Computer Science, Second Nordic Conference CAiSE'90, and Stockholm, Sweden. pp. 195-211.

Thirumal P. C. and Nagarajan N (2015). Utilization of Data mining techniques for Diagnosis of Diabetes Mellitus. A case study.

UNAIDS (2010). Global report: UNAIDS report on the global AIDS epidemic. Vol.2007, issue, Geneva.

UNAIDS (2011).World AIDS Day Report of 2011. Geneva UNAIDS. Retrieved From: www.unaids.org/.../unaids/.../unaidspublication/2011/JC2216 Accessed on March 2012.

USAID (2006). Bringing information to decision makers for global effectiveness: how HIV and AIDS affect population. Population Reference Bureau: Washington.

Vijendra and Shivani(2014).Robust Outlier Detection Technique in Data Mining: A Univariate Approach

Vincent H., 2007, 'Developing a Consumer Health Informatics Decision Support System Using Formal Concept Analysis'.

World Health Organization (WHO) (2010), Towards Universal Access: Scaling up priority HIV/AIDS interventions in the health sector. Progress Report, September 2010. Retrieved March 2012. URL: http://www.who.int/hiv/pub/2010progressreport/summary_en.pdf

World Health Organization (WHO). (2004) Policy statement on HIV testing. Retrieved on March, 2012. URL:http://www.who.int/hiv/pub/vct/statement/en/index.html

World Health Organization (WHOS) (2005).Scaling-up HIV testing and counseling services: a toolkit for program managers. Geneva, Switzerland.

World Health Organization. (2007). WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children.
Lu, H., Setiono, R., & Liu, H. (1996). Effective data mining using neural networks. *IEEE transactions on knowledge and data engineering*, *8*(6), 957-961.

Appendix

Appendix 1: Descriptions of Selected Attributes

SN	ATTRIBUTES	DESCRIPTION	Codes & Corresponding Values
1	ID	Person's Serial numbers	
2	Sex	Person's sex	{Male=M, Female=F}
3	Age	Person's age	{ less than 18, less than 18}
4	Test_In_labFile	Whether the person is tested or not	{ No=0,yes=1}
5	Final_testResult	Person's test result	{Negative=0, positive=1, Not-tested=2}
6	Level_of_Education	Person's education level	{No education=0primary=1, secondary=2, higher=3 }
7	marital status	Person's marital status	{never married=0,Married=1, divorced=2, widowed=3 }
8	NO of sex partner	Number of sexual partners a person have	{None=0, one=1, Two=2, More than two=3}
9	ever had sex	Whether the person ever had done sexual intercourse	{ No=0, Yes=1}
10	using condom during sex	Whether the person used condom during sexual intercourse	{ No=0,Yes=1}

Appendix 2: Sample Summary of Confusion Matrix used for Experimentation

A. Naivebayes algorithm with 10- folds cross validation.

Time taken to build model: 0.19 seconds

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

Correctly Classified Instances	14565	98.5053 %
Incorrectly Classified Instances	221	1.4947 %
Kappa statistic	0.9591	
Mean absolute error	0.0191	
Root mean squared error	0.0982	
Relative absolute error	7.6039 %	
Root relative squared error	27.7215 %	
Total Number of Instances	14786	

=== Detailed Accuracy By Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.061	0.981	1	0.99	0.98	Negative
1	0	1	1	1	1	Not_tested
0	0	0	0	0	0.751	Positive
Weighted Avg. 0.985	0.046	0.971	0.985	0.978	0.981	

=== Confusion Matrix ===

a	b	c	<-- classified as
11162	1	0 |	a = Negative
1	3403	0 |	b = Not_tested
219	0	0 |	c = Positive

B. J48 algorithm with 10- folds cross validation.

Time taken to build model: 0.54 seconds

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

Correctly Classified Instances 14566 98.5121 %
Incorrectly Classified Instances 220 1.4879 %
Kappa statistic 0.9593
Mean absolute error 0.0194
Root mean squared error 0.0986
Relative absolute error 7.7272 %
Root relative squared error 27.8088 %
Total Number of Instances 14786

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.061	0.981	1	0.99	0.97	Negative
1	0	1	1	1	1	Not_tested
0	0	0	0	0	0.615	Positive
Weighted Avg.	0.985	0.046	0.971	0.985	0.978	0.971

=== Confusion Matrix ===

a	b	c	<-- classified as
11163	0	0	a = Negative
1	3403	0	b = Not_tested
219	0	0	c = Positive

C. J48 algorithm with the percentage split.

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances 4359 98.2642 %
Incorrectly Classified Instances 77 1.7358 %
Kappa statistic 0.9535
Mean absolute error 0.0202
Root mean squared error 0.1064
Relative absolute error 7.9949 %
Root relative squared error 29.6712 %
Total Number of Instances 4436

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1	0.069	0.977	1	0.989	0.966	Negative
1	0	1	1	1	1	Not_tested
0	0	0	0	0	0.62	Positive
Weighted Avg.	0.983	0.051	0.966	0.983	0.974	0.968

=== Confusion Matrix ===

a	b	c	<-- classified as
3316	0	0	a = Negative
0	1043	0	b = Not_tested
77	0	0	c = Positive

D. MLP algorithm with 10- folds cross validation.

=== Confusion Matrix ===

a	b	c	<-- classified as
11162	1	0	a = Negative
1	3403	0	b = Not_tested
219	0	0	c = Positive

Appendix3. Sample values of the final selected attributes

ID	Sex	Age	Test_In_labFile	Final_testResult	Level_of_Education	marital status	NO of sex partner	ever had sex	using condom during sex
1	Male	more than 18	Yes	Negative	No_Education	never married	None	no	No
2	Female	more than 18	Yes	Negative	No_Education	married	One	yes	No
3	Male	more than 18	Yes	Negative	Primary_Education	widowed	Two	yes	yes
4	Female	more than 18	Yes	Negative	No_Education	married	One	yes	No
5	Male	more than 18	Yes	Negative	No_Education	divorced	more than two	yes	yes
6	Female	less than 18	no	Not_tested	Primary_Education	widowed	Two	yes	yes

Appendix4. Sample CSV (Comma Separated Value) data format

ID ,Sex,Age,Test_In_labFile,Final_testResult,Level_of_Education,martial status,NO of sex partner,ever had sex,using condom during sex

1,Male,more than 18,Yes,Negative,No_Education,never married,None,no,no

2,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

3,Male,more than 18,Yes,Negative,Primary_Education,widowed,two,yes,yes

4,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

5,Male,more than 18,Yes,Negative,No_Education,divorced,more than two,yes,yes

6,Female,les than 18,no ,Not_tested,Primary_Education,widowed,two,yes,yes

7,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

8,Female,more than 18,No,Not_tested,No_Education,divorced,more than two,yes,yes

9,Male,les than 18,Yes,Negative,Secondary_Education,married,one,yes,no

10,Male,more than 18,Yes,Negative,No_Education,divorced,more than two,yes,yes

11,Female,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

12,Female,les than 18,Yes,Negative,Secondary_Education,married,one,yes,no

13,Female,les than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

14,Female,more than 18,Yes,Negative,No_Education,never married,None,no,no

15,Male,les than 18,Yes,Negative,Primary_Education,married,one,yes,no

16,Female,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

17,Male,more than 18,Yes,Negative,No_Education,married,one,yes,no

18,Male,more than 18,Yes,Negative,No_Education,divorced,more than two,yes,yes

19,Female,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

20,Male,les than 18,Yes,Negative,Secondary_Education,married,one,yes,no

21,Female,les than 18,Yes,Negative,Primary_Education,divorced,more than two,yes,yes

22,Male,more than 18,Yes,Negative,Primary_Education,married,one,yes,no

23,Female,more than 18,Yes,Negative,No_Education,divorced,more than two,yes,yes

24,Male,more than 18,No,Not_tested,Primary_Education,never married,None,no,no

25,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

26,Male,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

27,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

28,Male,more than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

29,Female,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

30,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

31,Male,more than 18,Yes,Negative,No_Education,divorced,more than two,yes,yes

32,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

33,Male,more than 18,Yes,Positive ,Secondary_Education,divorced,more than two,yes,yes

34,Female,more than 18,Yes,Positive ,No_Education,widowed,two,yes,yes

35,Male,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

36,Male,more than 18,No,Not_tested,Higher_Education,divorced,more than two,yes,yes

37,Female,more than 18,Yes,Negative,Primary_Education,never married,None,no,no

38,Female,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

39,Female,more than 18,Yes,Negative,Secondary_Education,widowed,two,yes,yes

40,Male,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

41,Female,more than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

42,Male,more than 18,Yes,Negative,Secondary_Education,widowed,two,yes,yes

43,Male,les than 18,No,Not_tested,Secondary_Education,married,one,yes,no

44,Male,more than 18,No,Not_tested,Higher_Education,divorced,more than two,yes,yes

45,Female,more than 18,No,Not_tested,Secondary_Education,married,one,yes,no

46,Male,more than 18,No,Not_tested,Higher_Education,divorced,more than two,yes,yes

47,Female,les than 18,Yes,Negative,Secondary_Education,never married,None,no,no

48,Female,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

49,Male,more than 18,No,Not_tested,Primary_Education,widowed,two,yes,yes

50,Female,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

51,Female,more than 18,Yes,Negative,Primary_Education,divorced,more than two,yes,yes

52,Male,more than 18,No,Not_tested,Higher_Education,widowed,two,yes,yes

53,Male,more than 18,No,Not_tested,Primary_Education,married,one,yes,no

54,Female,more than 18,No,Not_tested,Higher_Education,divorced,more than two,yes,yes

55,Male,more than 18,Yes,Negative,Higher_Education,married,one,yes,no

56,Female,more than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

57,Male,more than 18,No,Not_tested,Higher_Education,widowed,two,yes,yes

58,Female,more than 18,Yes,Negative,Higher_Education,married,one,yes,no

59,Male,more than 18,Yes,Negative,Higher_Education,divorced,more than two,yes,yes

60,Female,more than 18,Yes,Negative,Higher_Education,never married,None,no,no

61,Female,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

62,Female,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

63,Female,more than 18,No,Not_tested,Higher_Education,married,one,yes,no

64,Male,more than 18,No,Not_tested,Secondary_Education,divorced,more than two,yes,yes

65,Female,les than 18,No,Not_tested,Secondary_Education,widowed,two,yes,yes

66,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

67,Male,more than 18,No,Not_tested,Higher_Education,divorced,more than two,yes,yes

68,Female,more than 18,No,Not_tested,Secondary_Education,married,one,yes,no

69,Female,more than 18,Yes,Negative,Higher_Education,divorced,more than two,yes,yes

70,Female,more than 18,Yes,Negative,Secondary_Education,never married,None,no,no

71,Male,more than 18,Yes,Negative,Higher_Education,married,one,yes,no

72,Female,more than 18,Yes,Negative,Secondary_Education,widowed,two,yes,yes

73,Female,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

74,Male,more than 18,No,Not_tested,Secondary_Education,divorced,more than two,yes,yes

75,Male,more than 18,No,Not_tested,Secondary_Education,widowed,two,yes,yes

76,Male,more than 18,No,Not_tested,Secondary_Education,married,one,yes,no

77,Female,more than 18,No,Not_tested,No_Education,divorced,more than two,yes,yes

78,Male,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

79,Male,more than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

80,Male,more than 18,Yes,Negative,Secondary_Education,widowed,two,yes,yes

81,Male,more than 18,No,Not_tested,Secondary_Education,married,one,yes,no

82,Female,more than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

83,Female,les than 18,Yes,Negative,No_Education,never married,None,no,no

84,Male,more than 18,Yes,Negative,No_Education,married,one,yes,no

85,Female,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

86,Male,more than 18,Yes,Negative,Secondary_Education,married,one,yes,no

87,Female,les than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

88,Female,more than 18,No,Not_tested,Secondary_Education,widowed,two,yes,yes

89,Male,more than 18,Yes,Negative,Higher_Education,married,one,yes,no

90,Female,more than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

91,Male,les than 18,Yes,Negative,Secondary_Education,married,one,yes,no

92,Male,les than 18,Yes,Negative,Secondary_Education,divorced,more than two,yes,yes

93,Male,more than 18,Yes,Negative,Primary_Education,never married,None,no,no

94,Female,more than 18,No,Not_tested,No_Education,married,one,yes,no

95,Male,more than 18,Yes,Negative,No_Education,widowed,two,yes,yes

96,Female,more than 18,Yes,Negative,No_Education,married,one,yes,no

97,Female,les than 18,No,Not_tested,No_Education,divorced,more than two,yes,y

Appendix5. Association rules

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.7 (10350 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 4

Size of set of large itemsets L(3): 1

Best rules found:

1. Final_testResult=Negative 11163 ==> Test_In_labFile=Yes 11162 conf:(1)

2. Final_testResult=Negative ever had sex=yes 10566 ==> Test_In_labFile=Yes 10565 conf:(1)

3. Test_In_labFile=Yes ever had sex=yes 10770 ==> Final_testResult=Negative 10565 conf:(0.98)

4. Test_In_labFile=Yes 11381 ==> Final_testResult=Negative 11162 conf:(0.98)

5. Final_testResult=Negative 11163 ==> ever had sex=yes 10566 conf:(0.95)

6. Test_In_labFile=Yes Final_testResult=Negative 11162 ==> ever had sex=yes 10565 conf:(0.95)

7. Final_testResult=Negative 11163 ==> Test_In_labFile=Yes ever had sex=yes 10565 conf:(0.95)

8. Test_In_labFile=Yes 11381 ==> ever had sex=yes 10770 conf:(0.95)

9. Age=more than 18 12791 ==> ever had sex=yes 12083 conf:(0.94)

10. Test_In_labFile=Yes 11381 ==> Final_testResult=Negative ever had sex=yes 10565 conf:(0.93)