ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MINING ART DATA SET TO PREDICT CD4 CELLS COUNT THE CASE OF JIMMA, BONGA AND AMAN HOSPITALS

MISGANAW TADESSE

JUNE, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MINING ART DATA SET TO PREDICT CD4 CELLS COUNT THE CASE OF JIMMA, BONGA AND AMAN HOSPITALS

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULLFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN HEALTH INFORMATICS

BY

MISGANAW TADESSE

JUNE, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MINING ART DATA SET TO PREDICT CD4 CELLS COUNT THE CASE OF JIMMA, BONGA AND AMAN HOSPITALS

BY

MISGANAW TADESSE

Members of the examining board:

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

DECLARATION

I declare that this thesis is my original work and it has not been presented for a degree in any other university. All the material sources used in this work are duly acknowledged.

_____

Misganaw Tadesse

June, 2013

This thesis has been submitted for examination with our approval as university advisors.

_____

Ato Ermias Abebe

_____

Dr. Nigussie Deyasa

DEDICATION

I dedicate this work to my Mom and Dad who are always insisting me to continue my education to the level possible.

This dedication also goes to my adorable baby boy Nahom and to my beloved wife Kidist for their continued patience and love during the times I spent away from them to complete this work.

ACKNOWLEDGMENT

In the very first place, I would like to glorify the almighty GOD for giving me the chance, courage, diligence, patience and most importantly for sending me the blessings to have a positive feeling towards completing this thesis.

Second and most importantly, I would like to acknowledge my advisor Ato Ermias Abebe from school of Information Science for his wholehearted, dynamic and just in time comments and suggestions in every aspect of this work. Without his advice, the completion of this research would have been impossible.

My deepest gratitude also goes to Dr. Million Meshesha from school of Information Science for his support and comments on every of the problems I faced in this work.

My sincere appreciation also goes to those people who are working in the ART clinics of Jimma, Bonga and Aman Hospitals in counseling and record keeping. Especially data clerks Genet and Ayantu from Jimma, Deraritu from Bonga, and Simegn and Fitsum from Aman; their support was really wonderful.

I would like to acknowledge Eleni Seyoum from FHAPCO, for the precious comments I got to validate the findings of this work which made the work to be complete.

I would like to thank my class mates for those unforgettable best moments I had from the very onset of this program until the end. Especially Teketel Mulugeta, Dawit Ayele, Samson Kiflom, Hiwot Abebe, Wondwosen Shiferaw, Mekdes Tamiru and all health informatics graduates of 2012/13; you guys are really incredible.

Finally, I would like to thank the school of Information Science and Public health for setting the fund to run this study and still very much importantly the health informatics program coordinator Meseret Ayanaw for her continuous and timely notification of tasks to be completed in the specified time frame.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AIDS	Acquired Immunodeficiency Syndrome
ANN	Artificial Neural Network
ARFF	Attribute Relation File Format
ART	Antiretroviral Therapy
ART-LINC	Antiretroviral Therapy in Low Income Countries
ARV	Antiretroviral
AUC	Area under ROC Curve
CCI	Correctly Classified Instances
CD4	Clustered Differentiation level 4
CRISP	Cross Industry Standard Process
CSV	Comma Separated Value
DART	Development of Antiretroviral Therapy (in Africa)
DM	Data Mining
FDA	Food and Drug Administration
FHAPCO	Federal HIV/AIDS Prevention and Control Office
FMOH	Federal Ministry of Health
FPR	False Positive Rate
GUI	Graphical User Interface
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunity Virus
ICAP	International Center for AIDS Care and Treatment Programs
ICI	Incorrectly Classified Instances
IN	In care for other treatments
IQR	Inter Quartile Range
JHU-TSEHAI	Johns Hopkins University, Technical Support for Ethiopia HIV/AIDS ART Initiative
KDD	Knowledge Discover in Databases
KDP	Knowledge Discovery Process
MLP	Multilayer Perceptron
OART	On Antiretroviral Therapy

PART	Partial Decision Tree
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics
SEMMA	Sample Explore Modify Model Asses
SMO	Sequential Minimal Optimization
SPSS	Statistical Package for Social Science
SQL	Structured Query Language
SVM	Support Vector Machine
TPR	True Positive Rate
UNAIDS	Join United Nations program on HIV/AIDS
WEKA	Waikato Environment for Knowledge Learning
WHO	World Health Organization

ABSTRACT

Background: Recent reports from WHO and UNAIDS indicate that the number of people using ART are increasing from time to time. This number is dramatically increasing in sub Saharan African countries including Ethiopia. According to the report of WHO and UNAIDS, as of the end of 2011, over 8 million people had access to ART in low and middle-income countries.

Objective: The purpose of this study is to apply data mining techniques on ART records of patients maintained in Jimma, Bonga and Aman Hospitals ART database to build a model capable of predicting CD4 cells count of patients after six, twelve and eighteen months of treatment.

Methodology: The overall activity of this thesis is guided by a Hybrid-DM model which is a six step knowledge discovery process model. The study has used 7,252 instances, ten predicting and three outcome variables to run the experiments. Due to the nature of the problem and attributes contained in the dataset, classification mining task is selected to build the classifier models. The mining algorithms; J48, PART, SMO and MLP are used in all experiments due to their popularity in recent related works. In addition to base classifiers, due to the imbalanced nature of classes in each of the three outcome variables, a boosting algorithm (AdaBoostM1) is used to boost the classifiers predictive performance. Ten-fold cross validation technique is used to train and test the classifier models. Performance of the models is compared using accuracy, TPR, FPR, mean absolute error, F-measure, and the area under the ROC curve.

Results: The boosting algorithm has given the base classifier a better predictive accuracy with the PART unprunned decision tree yielding a better model of the sixth and twelfth month CD4 count, and the pruned PART decision tree performed better for the eighteenth month CD4 count. The joined rules of the three models indicated that, baseline CD4 count, drug-regimen, age, family planning usage status, WHO clinical stage, and functional status of a patient are the most determinant attributes used to predict CD4 counts.

Conclusion: A promising result is observed in applying data mining techniques to build CD4 count predictive model using socio-demographic, clinical and biological features. Future works can be done both on validating the results using clinical trials and also by doing the same study changing the source data or knowledge discovery techniques used in this work.

CHAPTER ONE

INTRODUCTION

1.1. Background

A decade ago, having AIDS was almost equivalent to a death sentence. Since 1996, with the introduction of combined antiretroviral treatment, AIDS has become chronic, manageable disease [1].

ART is treatment of people infected with human immunodeficiency virus using anti-HIV drugs [38]. The standard treatment consists of a combination of at least three drugs (often called HAART) that suppress HIV replication [38]. Three drugs are used in order to reduce the likelihood of the virus developing resistance. ART has the potential both to reduce mortality and morbidity rates among HIV-infected people, and thereby improve their quality of life [38].

The primary goal of ART is to reduce HIV-associated morbidity and mortality. This goal is best accomplished by using effective ART to maximally inhibit HIV replication, as defined by achieving and maintaining plasma HIV RNA (viral load) below levels detectable by commercially available assays [2]. Durable viral suppression improves immune function and quality of life, lowers the risk of both AIDS-defining and non-AIDS-defining complications, and prolongs life [2]. Based on emerging evidence, additional benefits of ART include a reduction in HIV-associated inflammation and possibly its associated complications [2].

While ART significantly decreases mortality, the latter is higher in the first six months than during the subsequent time on therapy, particularly when patients start with stage 4 clinical events, severe immune suppression and very low CD4 counts [3]. The ART-LINC collaboration (eighteen treatment programs in Africa, Asia and South America) recorded a 4% mortality rate in 2725 patients under active follow-up six months after starting therapy but noted that mortality fell to 2% in the subsequent six months of therapy [3]. The DART trial reported that 39 of 62 deaths (63%) in a cohort of over 1000 adults followed for two years occurred in the first six months of therapy [3].

At the end of 2010, an estimated 34 million people [31.6 million–35.2 million] were living with HIV worldwide, up 17% from 2001 [4]. This reflects the continued large number of new HIV

infections and a significant expansion of access to antiretroviral therapy, which has helped to reduce AIDS-related deaths, especially in more recent years.

Ethiopia is the second most populous and one of the seriously affected countries in sub Saharan Africa [5]. With more than 1.3 million people living with HIV and an estimated 277,800 people requiring treatment; the Government of Ethiopia has taken measures to reduce the risk of HIV transmission and mitigate the impact of the epidemic on society [5].

In 2010, WHO and UNAIDS launched the Treatment 2.0 strategy, which promotes radical simplification of ART, with accelerated treatment scale-up and full integration with prevention, in order to reach Universal Access [6]. Besides the previously developed treatment strategies, WHO is working to release a revised guideline on the use of antiretroviral in 2013, which will include recommendations on ART for adults and adolescents [6].

The most dramatic increases in antiretroviral therapy coverage have occurred in sub-Saharan Africa, with a 20% increase between 2009 and 2010 alone [4]. It is estimated that at least 6.6 million people in low and middle-income countries are receiving HIV treatment [2]. This is an increase of more than 1.35 million over the previous year [4]. In low and middle-income countries 47% of the 14.2 million eligible people living with HIV were on antiretroviral therapy at the end of 2010, compared to 39% at the end of 2009 [4].

According to the report of WHO and UNAIDS, as of the end of 2011, over 8 million people had access to ART in low and middle-income countries [6]. WHO is providing countries with ongoing guidance, tools and support in delivering and scaling up ART within a public health approach [6].

From the reports cited here and elsewhere, one can easily note that the number of patients participating in ART programs is increasing from time to time. Yet, the data maintained is rarely used in forecasting different scenarios where problems in the area can be solved. Hence, the records in ART databases contain many attributes which are representative of a given patient. Thus, by applying data mining techniques it is very possible to build a model of the essential patterns found within the ART dataset in order to predict progressive CD4 counts of patients taken within the range of six months by taking their socio-demographic, clinical and biological features.

1.2. Statement of the Problem

The CD4 cell count remains the strongest predictor of HIV related complications, even after the initiation of therapy [7]. The baseline pretreatment value is informative: lower CD4 counts are associated with smaller and slower improvements in counts [7]. However, precise thresholds that define treatment failure in patients starting at various CD4 levels are not yet established [7]. As a general rule, new and progressive severe immunodeficiency is demonstrated by declining longitudinal CD4 cell counts which should trigger a switch in therapy [7].

A study conducted by Mellors et al [8] has showed the relevance of measuring the CD4 cells count of ART taking patients regularly (within six months interval) in the following way:

> *'Patients starting with low CD4 counts may demonstrate slow recovery, but persistent levels below 100 cells/mm3 represent significant risk for HIV disease progression. Caution has to be noted are that inter current infections can result in transient CD4 count decreases, and that, with relatively infrequent monitoring (e.g. every six months), the true peak of the CD4 cell count may be missed ... The CD4 cell count can also be used to determine when not to switch therapy... In general, switching should not be recommended if the CD4 cell count is above 200 cells/mm3.'*

There are problems which are persistently occurring in those ART service providing hospitals in the course of taking the CD4 cells count of patients. According to the discussion made with physicians:

> *"Patients are not willing to come and take CD4 test regularly ... So, this is a very good opportunity for the development of drug resistance, co-infection with other diseases and even in the spread of drug resistant virus strains to other people."*

In addition to the frustration of patients, physicians also pointed out the complexity and time taking of the investigation in terms of resource exploitation and physician's time in the following way:

> *"Even though CD4 count is essential to monitor disease progression, still measuring the CD4 count of an HIV patient demands a long period of laboratory investigation (Estimated to be Five to seven hours) and sophisticated identification techniques which may end up unsuccessfully."*

This again is not feasible in countries like Ethiopia with a poor resource setting and unbalanced number of skilled professionals in the area.

The other main problem associated with CD4 count is, frequent failure happening on the CD4 counting machine which creates a great challenge in taking CD4 counts regularly in the scheduled time. Key informants told the researcher that, this problem is more frequent in the case of Bonga and Aman Hospitals where patients are frequently advised to go to Jimma in order to get their CD4 count results.

Relatively similar study was done by Behailu [9], which is on predicting CD4 status of patients. In that study the CD4 status of patients during initiation of ART was modeled by taking socio-demographic attributes, ART start date, WHO clinical stage, Previous ART status, Drug regimen, ART status, Functional Status and Eligibility reason to build the model. The data items were classified according to the predefined classes named "**Low**" and "**Normal**" which were labels given to Baseline CD4 counts below and above the cutoff point **200** respectively. As a result, the following major problems were identified from that study.

The first problem was associated with taking of a baseline CD4 count as a target variable of prediction. But this variable is not actually a count taken after the initiation of treatment, rather it is a count taken when a patient is eligible for ART.

The second problem was, since CD4 count varies irregularly with time and other factors, it would have been better if two or three time counts after the initiation of therapy were taken to know the actual peak of CD4 count as Mellors et al [8] suggests. But here no follow up CD4 count data was used to develop the stated model.

The third problem comes in terms of the data mining techniques used; Behailu [9] has only incorporated J48 decision tree classifier and PART rule induction algorithm to develop the model. In a study conducted by Yashik et al [10] on forecasting CD4 count change, a support vector machine is used to model changes in CD4 count with a classification accuracy of 83%. Another study by Wang et al which is cited on Yashik et al [10] has confirmed the use of artificial neural network to build a model capable of predicting the viral load of HIV patients with an accuracy of 75%. This is a good evidence to use the two mentioned data mining techniques in addition to the previous ones.

The last problem was, findings of the study did not reach to the people experiencing the problem, but rather a model of the identified pattern was developed and left on the paper.

This indicates the study has left a very strong gap which has to be worked out to come up with a model that can predict CD4 count of ART following patients at sixth, twelfth and eighteenth months of therapy.

Taking the above problems into consideration, this research is primarily intended to develop a model which can predict CD4 cells count of ART following patients by taking ART records from Jimma, Bonga and Aman Hospitals. This will save a great deal of time and it adds value to the quality of decisions made by physicians. On the other hand, it helps the patient to have improved quality of life and most importantly it can be used in facilities where there is a shortage of professionals or lack or problem with CD4 counting machines.

To this end, this research attempts to answer the following research questions:

- Which classification algorithm is more suitable to build CD4 cells count predictive model?
- Which attributes are more important to predict the CD4 cells count of ART following patients?

1.3. Objectives of the Study

1.3.1. General objective

The main objective of this study is to apply data mining techniques on ART records of patients maintained in Jimma, Bonga and Aman Hospitals ART database to build a model capable of predicting CD4 cells count of patients after six, twelve and eighteen months of treatment.

1.3.2. Specific Objectives

The specific objectives are to:

- Prepare the data for analysis and model building by cleaning, extracting, and transforming in to a format suitable for the data mining algorithm.
- Apply classification algorithms to train, test and build classifier models.

- Compare and select the best model using parameters; true positive rate, false positive rate, F-measure, accuracy, and area under the ROC curve (AUC).
- Identify more important attributes used in predicting CD4 cells count.
- Develop a graphically user friendly prototype system (interface) of the model to ease usage of the knowledgebase by domain users.
- Report the research findings and make recommendations.

1.4. Significance of the study

The findings of this research could be used to predict the CD4 count of patients attending ART programs after six, twelve and eighteen months of treatment taking the ten predicting variables suggested by this study. This has a great deal of benefit to patients, physicians, hospitals, policy makers, researchers and local and international partners working on supporting the implementation of ART programs.

Accordingly, the benefits offered to these parties are:

- Physicians can make use of the model to forecast CD4 counts of their patients ahead which in turn eliminates complications which might occur due to not knowing the CD4 count during the right time. In addition to this, in facilities with no CD4 counting machine it can be used as a replacement.
- Patients receive the right regimen associated to their CD4 level so that their immunity level develops to a level where no opportunistic infection can occur. The waiting time to get the CD4 count is not going to be a problem for the patients.
- Policy makers can make use of the model to develop new guidelines or modify the existing one in order to improve implementation of ART programs in the country.
- The result obtained in this study can be taken as a base to conduct clinical investigation to validate the findings with the real situations and also similar researches can be done in some other part of the country to validate the findings and then make it to serve at a national level.
- Local and international partners can make use of the findings to identify the kind of support expected from them.

1.5. Scope and Limitations of the study

This study is limited to predicting CD4 cells count of patients based on useful patterns or relationships found in ART records kept in databases of Jimma, Bonga and Aman Hospitals. This makes the finding to be limited to that specific region.

The first exclusion criterion is accompanied with the enrollment year of the patients in to the program. Therefore, the study dataset records comprise those which are only recorded from 1996 E.C to 2004 E.C. Thus, those which are found either above or below this two demarcation years are discarded from the study dataset.

The second exclusion criteria is the kind of patient records to be included in the study; only patient's records with the ART status of "**OART**" i.e. only those who are started taking the drug are eligible whereas others like "**IN**" i.e. in care for other treatments are rejected from the dataset.

The third exclusion criterion was the age of the patient's. For this attribute those records having a value of eighteen (18) and above were included whereas those below this age value are discarded. This has happened due to two reasons. The first one is, naturally CD4 count decreases with age, so those with age of 18 and above has relatively approaching CD4 count. The other reason is almost 90% of instances in the dataset have the age value above 18 years and the remaining 10% are below 5 years of age. The CD4 count of children is taken in percentage form and that of adults is absolute count so that it is difficult to treat these two groups of subjects together. Therefore, those who are below 18 years are removed from the study subjects.

The Fourth exclusion criteria used was the CD4 counts taken at the baseline, after six, twelve and eighteenth months. Thus, those patient's records missing at least one of these CD4 counts was discarded from the dataset.

Regarding the data mining tasks, classification mining techniques were used by initially creating groups of CD4 cells count ranges for the three dependent variables. Among the available classification algorithms, decision tree (J48), rule induction (PART), artificial neural network (MLP), and support vector machine (SMO) were used to build the models.

The study is limited to developing a predictive model from attributes found in the ART dataset. But these are not enough to make an approximate prediction of CD4 count. There exist variables

like feeding style, BMI, drug adherence, drug abuse, economical status, and others which might have a role in determining one's CD4 status. Therefore, missing such variables within the study dataset has an effect on the accuracy of the findings.

1.6. Ethical Consideration

The researcher has taken in to consideration three major ethical issues in the course of the study. The first one is the research does not use personal identifiers like name during model building or in reporting its findings. The second one is the findings of this work are not to be used for commercial gain. Moreover, the findings of the study would not harm study subjects. Ethical clearance was obtained from School of Public health and a letter of cooperation was obtained from school of Information Science before the launch of data collection.

1.7. Organization of the Thesis

This thesis is organized in to six chapters. The first chapter deals with the general overview of the study including background, statement of the problem, research objectives, significance of the research, scope and limitations of the research.

The second chapter focuses on literature review on data mining technology, health care and data mining, HIV/AIDS and ART, and also extensive review of related works are included.

Chapter Three has included two main topics of this study. The first one is about data mining algorithms used to build the model and the second one is the research methodology incorporated to guide the entire research work.

The fourth chapter is about data preparation which constitutes business understanding, data understanding and data preprocessing. Therefore, at this stage of modeling a quality data is made ready for the classification algorithms.

Chapter Five is where the experiments conducted are presented. Here, topics about data mining, model selection and Prototype development were discussed in detail. All experiments have passed three stages of model development i.e. training, building and validation of the models. Results of the experiments are also analyzed and interpreted.

Chapter six is the final chapter which presents summary, concluding remarks and recommendations of the study.

CHAPTER TWO

LITERATURE REVIEW

Nowadays data mining is becoming widely used in every field (be it in health, agriculture, education, business, etc.) for extracting hidden knowledge which is crucial for competitive advantage and sustainable growth. In this chapter, a thorough investigation has been made to survey and incorporate a range of standards and steps used in data mining methodologies, various concepts, theories and practices of data mining in relation with issues related to HIV/AIDS and ART. Moreover, related research works which reveal the applicability of data mining technology on HIV/AIDS records in general and ART records specifically are reviewed.

2.1. Conceptual Review

2.1.1. Data Mining

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 [11]. Data mining can be considered a relatively recently developed methodology and technology coming into prominence only in 1994 [12].

One of the aims of data mining can be seen as the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful. This relationships and summaries derived through data mining are often referred to as models or patterns. A model is a high-level description, summarizing a large collection of data and describing its important features. The structure of the model is a global summary of a dataset. In contrast to the global nature of models, local patterns make statements only about restricted regions of the space spanned by the variables [13].

The patterns discovered during data mining must be meaningful in that they usually lead to an economic benefit. Researchers often strive to discover the patterns that govern how the physical world works and encapsulate them in theories that can be used for predicting what will happen in new situations [14].

2.1.2. Common Methodologies in Data Mining

In the latest years, it has been seen that data mining area is growing at a phenomenal rate. Some efforts are being done that seek the establishment of standards in the area, both by academics and by people in the industry field. The academics efforts are centered in the attempt to formulate a general framework for DM [15]. The bulk of these efforts are centered in the definition of a language for DM that can be accepted as a standard, in the same way that SQL was accepted as a standard for relational databases [16, 17, 18, 19, 20]. The efforts in the industrial field concern mainly the definition of processes/methodologies that can guide the implementation of DM applications. In this paper, KDD, SEMMA, CRISP-DM and Hybrid-DM are compared to select the most suitable modeling for this study.

2.1.2.1. KDD

The KDD process, as presented in Fayyad et al [11], it is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. There are considered five stages, presented in figure 2.1 which are taken from Fayyad et al [11] to clearly indicate the processes involved in knowledge discovery.

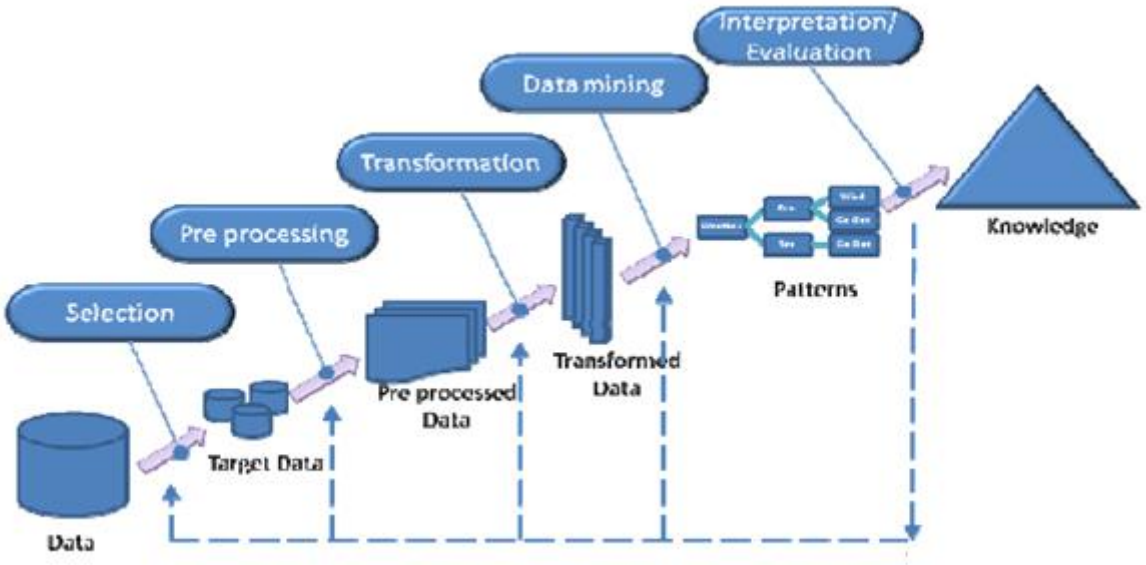

Figure 2.1: KDD- Process Model

As figure 2.1 shows several steps from selecting data to providing understandable knowledge to the user are involved in the KDD process. According to Fayyad et al [11] the details of the explanation on the KDD process model is explained in the following steps.

Step 1. Selection: Before directly going to selection one needs to understand the domain and identify the goal of applying the process. Then select the necessary data to solve the problem. The main part of the KDD process is the selection of raw data that is necessary for the discovery. There might be unnecessary data attributes provided, but only few data attributes are needed in the process. Selecting the necessary data attributes for the discovery places an important part which yields target dataset.

Step 2. Preprocessing: Handling missing values, eliminating noise and duplicate records in the data sets is the main part of this process. Missing values in the data lead to loss of useful information, which might not result in discovering useful knowledge. Noise in data and duplicate records mislead the process in obtaining accurate knowledge. Therefore, data cleaning and the preprocessing are necessary to produce better quality results. There are various tools such as Microsoft Excel, SPSS and Weka used for these tasks.

Step 3. Transformation: In this step data is transformed or consolidated into forms appropriate for data mining. Data transformation involves the following. First smoothing of data, where the noise from data is removed; second aggregation of data, where aggregation operations are applied to data for the analysis of the data; third, generalization, where raw data is replaced with high level concepts and finally normalization, where the attribute data are scaled to fall within a specified range; and feature Selection, where new attributes are constructed and added from the given set of attributes.

Step 4. Data mining: The step that requires analysis of the main problem and decision on which models and parameters are appropriate. Depending on the model, different data mining algorithms and methods are chosen that are needed for searching data patterns. Data mining methods are performed to achieve the goal by finding the interesting patterns in the data. Better results are obtained if the preceding steps are performed properly.

Step 5. Interpretation/Evaluation: The step where the mining results are interpreted and even the process can start again from step 1 if there are any errors or for providing further accurate results. It involves the visualization of extracted patterns and results. The knowledge discovery

process then takes the raw results from data mining and transforms them into useful and understandable information for the users.

Using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties, checking for and resolving potential conflicts with previously believed or extracted knowledge can be part of this step.

2.1.2.2. CRISP-DM

David L and Dursen D explained that Cross-Industry Standard Process for Data Mining (CRISP-DM) is a knowledge discovery approach which is widely used by industry members [21]. This model consists of six phases intended as a cyclical process (see **Figure 2.2**). According to David L and Dursen D all the steps included in this model are explained in the paragraphs below [21].

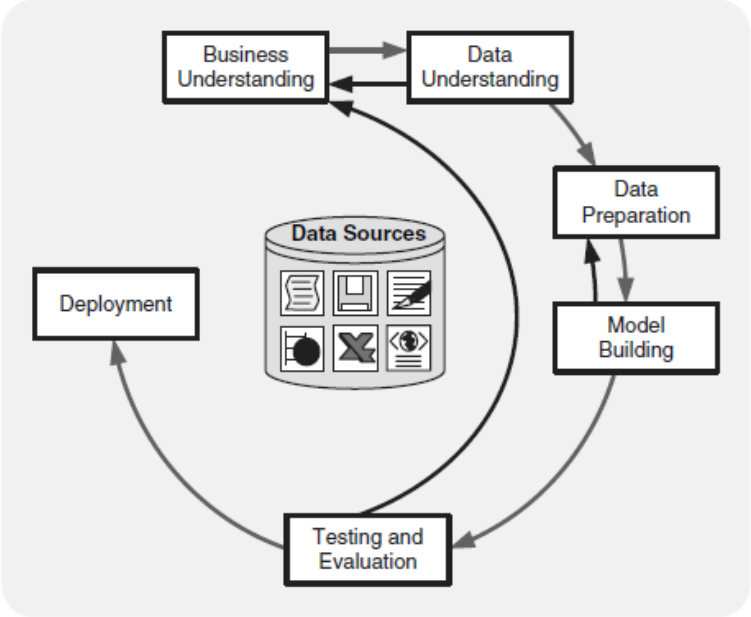

Figure 2.2: CRISP-DM Process model

Business Understanding: Business understanding includes determining business objectives, assessing the current situation, establishing data mining goals, and developing a project plan.

Data Understanding: Once business objectives and the project plan are established, data understanding considers data requirements. This step can include initial data collection, data description, data exploration, and the verification of data quality. Data exploration such as viewing summary statistics (which includes the visual display of categorical

variables) can occur at the end of this phase. Models such as cluster analysis can also be applied during this phase, with the intent of identifying patterns in the data.

Data Preparation: Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. Data cleaning and data transformation in the preparation of data for modeling needs to occur in this phase. Data exploration at a greater depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding.

Modeling: Data mining software tools such as visualization (plotting data and establishing relationships) and cluster analysis (to identify which variables go well together) are useful for initial analysis. Tools such as generalized rule induction can develop initial association rules. Once greater data understanding is gained, more detailed models which are appropriate to the data type can be applied. The division of data into training and test sets is also needed for modeling.

Evaluation: Model results should be evaluated in the context of the business objectives established in the first phase (business understanding). This will lead to the identification of other needs (often through pattern recognition), frequently reverting to prior phases of CRISP-DM. Gaining business understanding is an iterative procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations.

Deployment: Data mining can be used to both verify previously held hypotheses, or for knowledge discovery (identification of unexpected and useful relationships). Through the knowledge discovered in the earlier phases of the CRISP-DM process, sound models can be obtained that may then be applied to business operations for many purposes, including prediction or identification of key situations. These models need to be monitored for changes in operating conditions, because what might be true today may not be true a year from now. If significant changes do occur, the model should be redone. It's also wise to record the results of data mining projects so documented evidence is available for future studies.

This six-phase process is not a rigid by-the-numbers procedure [21] rather there is a great deal of backtracking in between the phases. In addition to this, experienced analysts may not need to apply each phase for every study [21].

2.1.2.3. SEMMA

The SEMMA process was developed by the SAS Institute [22]. 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 five stages for the process [22]:

1. **Sample** – This stage consists of 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 of the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

3. **Modify** – This stage consists of the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

4. **Model** – This stage consists of 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 of assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

Although the SEMMA process is independent from the DM chosen tool, it is linked to the SAS Enterprise Miner software and pretends to guide the user on the implementations of DM applications [23]. SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for this conception, creation and evolution, helping to present solutions to business problems as well as to find the DM business goals [23].

2.1.2.4. Hybrid-DM

Hybrid-DM model is developed by Cios et al based on the CRISP-DM model by adopting it to academic researches [24]. It is a six stage process modeling which constitutes; Understanding of

the Problem, Understanding of the Data, Preparation of Data, Data Mining, Evaluation of the Discovered Knowledge and Use of the Discovered Knowledge.

2.1.3. Assessment of Data Mining Methodologies

The comparison among the four methodologies is done based on the steps contained, applicability to academic researches in general and its applicability to the identified problem.

The first comparison is done by comparing KDD and SEMMA data mining methodologies in the steps contained to govern the data mining task. By doing a comparison of the KDD and SEMMA stages we would affirm that they are equivalent.

- Sample can be identified with Selection,
- Explore can be identified with Pre processing
- Modify can be identified with Transformation
- Model can be identified with Data Mining
- Assess can be identified with Interpretation/Evaluation.

Examining it thoroughly, we may affirm that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process, since it is directly linked to the SAS Enterprise Miner software. But the problem with both SEMMA and KDD is; there is no stage which can let the researcher to get acquainted with the business in order to have initial insights about the problem and a stage to communicate the findings with the domain users.

Comparing the KDD stages with the CRISP-DM stages is not as straightforward as in the SEMMA situation. Nevertheless, we can first of all observe that the CRISP-DM methodology incorporates the steps that as referred above must precede and follow the KDD process, that is to say:

- The Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user
- The Deployment phase can be indentified with the consolidation by incorporating this knowledge into the system.

Concerning the remaining stages;

- The Data Understanding phase can be identified as the combination of Selection and Pre processing
- The Data Preparation phase can be identified with Transformation
- The Modeling phase can be identified with Data Mining
- The Evaluation phase can be identified with Interpretation/Evaluation.

This shows that CRISP is better than KDD as well as SEMMA by including additional stages for business understanding and deploying the final outcome or discovered knowledge in to the existing system.

When we look at CRISP-DM and Hybrid-DM they are almost equivalent with the development stages contained but with some adjustments done for Hybrid-DM. The main differences and extensions with CRISP-DM are, hybrid-dm:

- Provides more general, research-oriented description of the steps,
- Introduced a data mining step instead of the modeling step,
- Introduced 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.

Considering the above details on both CRISP and Hybrid-dm process modeling, the one which suits this research is Hybrid-DM modeling due to its nature to be made for academic environments and the opportunity of using the discovered knowledge in to other domains.

The other important reason for the selection of hybrid-dm modeling is; it is initially tested with many health related works.

The summary of the above comparison is presented in a table below.

KDD	SEMMA	CRISP-DM	Hybrid
Pre KDD	----------	Business Understanding	Understanding of the problem
Selection	Sample	Data Understanding	Understanding of the data
Preprocessing	Explore		
Transformation	Modify	Data Preparation	Preparation of the data
Data mining	Model	Modeling	Data Mining
Interpretation/Evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
Post KDD	----------	Deployment	Use of the Discovered Knowledge

Table 2.1: Summary of the Correspondences between KDD, SEMMA, CRISP-DM and Hybrid-DM

2.1.4. Data Mining Functionalities and Associated Algorithms

As noted on Han et al [25], data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. The detail about the major data mining functionalities is explained in the following paragraphs.

2.1.4.1. Characterization and Discrimination

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. According to Han et al [25], these concepts and class descriptions can be derived using data characterization and data discrimination. Data characterization is a summarization of the general characteristics of a target class of data and data discrimination is a comparison of the general features of a target class data objects with the general features of objects from one or a set of contrasting classes [25].

2.1.4.2. Mining Frequent Patterns, Associations and Correlations

Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including itemsets, subsequences, and substructures [25]. As

noted by Han et al [25], a frequent **itemset** typically refers to a set of items that frequently appear together in a transactional data set. A frequently occurring subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern [25]. A substructure can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets or subsequences [25]. If a substructure occurs frequently, it is called a (frequent) structured pattern [25]. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

2.1.4.3. Classification and Prediction

As noted by Han et al [25], classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e. data objects whose class label is known).

The derived model can be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks [25]. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [25]. Decision trees can easily be converted to classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units [25]. There are many other methods for constructing classification models, such as naïve bayesian classification, support vector machines, and *k*-nearest neighbor classification [25].

Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions [25]. That is, it is used to predict missing or unavailable numerical data values rather than class labels. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well [25].

2.1.4.4. Cluster Analysis

As described by Han et al [25], unlike classification, in clustering class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity [25]. That is, clusters of objects

are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [25]. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that groups similar events together [25].

2.1.5. Selected Mining Technique

Hence the intention of this paper is predicting the CD4 count of patients who have started taking antiretroviral treatment in ART clinics. The prediction is based on the socio-demographic data, clinical data and CD4 counts taken during initiation of ART. The outcome variables predicted using the above independent attributes are CD4 count after six months, CD4 count after twelve months and CD4 count after eighteen months. The prediction is made by splitting the CD4 counts in to ranges and assigning a unique label for each of these ranges of CD4 cells count. To achieve this target a supervised learning data mining task (classification) is better than the rest to address the study objectives.

Every classification method has its own strengths and limitations and that real world problems do not always satisfy the assumptions of a particular method, one approach is to apply all appropriate methods and select the one that provides the best solution [44]. Accordingly this paper has employed Decision tree, Rule Induction, Artificial Neural Network and Support vector machine classification techniques to build and test classifiers. These methods are selected due to their frequent appearance in recently published and unpublished researches [9, 10, 53].

The details about each of the above mentioned data mining techniques and accompanied algorithms is deeply covered in the third chapter, **Section 3.1**.

2.1.6. Data mining Tools Evaluation

Experience shows that no single machine learning scheme is appropriate to all data mining problems [26]. The development and application of data mining algorithms requires the use of powerful software tools. Nowadays it is very often possible to look for many kinds of data mining tools both commercially and freely. In this paper the main concern regarding mining tools focuses on free or General Public License (GPL) mining tools due to the availability and cost factors faced by the researcher on commercial mining software packages. Therefore, the

comparison is made only on free data mining software tools while excluding commercial ones. Still problem roses on the selection of the right tool due to the availability of large number of such tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult [27]. In this paper the researcher has tried to compare five well known open source mining software packages namely Rapid Miner, Weka, Orange, Rattle and KNIME as presented below.

Weka, RapidMiner and KNIME are developed in the Java software language [28]. Rattle is a fully R-based application, and Orange is integrated with Python [28]. Among all open source data mining tools, Weka and RapidMiner have the biggest and most active user communities [28]. Both of them quickly implement (and integrate) new and emerging machine learning algorithms into their systems [28].

Weka, as one of the best-known open-source data mining software tool it has an impressive array of data mining components which have, in fact been integrated into many other data mining tools including RapidMiner, Rattle, and KNIME [28]. Weka consists of four major applications: Explorer (for exploring data), Experimenter (for performing experiments and conducting statistical tests between learning schemes), Knowledge Flow (for incremental learning), and Simple CLI (a command-line interface to allow direct execution of WEKA commands) [28].

Rapider Miner (formerly YALE) is built on top of Weka, and includes additional powerful data analysis functions such as data preprocessing, visualization, and additional machine learning algorithms. In addition, its user interface it is more intuitive than Weka Knowledge Flow [28]. Moreover Rapid Miner implements the full Weka catalog as well as its own library to process the data mining problems.

KNIME has one of the best built-in on-line support features, which is very helpful for new users who are in the process of building their data mining workflows. KNIME also supports running R and Python scripts [28].

Rattle provides a GUI specifically for data mining using R [28]. Although an understanding of R is not required to begin using Rattle for basic data mining functions, Rattle is particularly suited for users familiar with R [28]. In addition, Rattle integrates two sophisticated tools for interactive graphical data analysis: GGobi and Latticist [28].

Orange has a very simple and intuitive graphical interface (GUI) for users with limited knowledge in data mining [28]. Compared to the other data mining tools, its strength is its interactive visualization function, which enables users to set visualization parameters and choose data points or nodes directly from a graph [28].

The detail of the comparison among these open source mining software packages is presented in the table below based on data mining functionalities, user interface and system features they accommodate [28].

Functionality	**Rapid Miner (YALE)**	**WEKA**	**Orange**	**Rattle**	**KNIME**
Bayes Network	Yes	Yes	Yes	No	Yes
Decision tree	Yes	Yes	Yes	Yes	Yes
Neural network	Yes	Yes	No	No	Yes
SVM	Yes	Yes	Yes	Yes	Yes
Feature selection	Yes	Yes	No	No	Yes
Clustering	Yes	Yes	Yes	Yes	Yes
Association rules	Yes	Yes	Yes	Yes	Yes
Model Information	Yes	Yes	Yes	Yes	Yes
Evaluation	Yes	Yes	Yes	Yes	Yes

Table 2.2: Comparison of Data Mining tools in terms of Mining Functionality

Mining Tool	**Rapid Miner (YALE)**	**WEKA**	**Orange**	**Rattle**	**KNIME**
Availability of User Interface	Yes	Yes	Yes	Yes	Yes

Table 2.3: Comparison of Data Mining tools in terms of User Interface

System Features	Rapid Miner (YALE)	WEKA	Orange	Rattle	KNIME
OS Platform	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux	Windows, Mac OS X, Linux
Documentation	4	5	3	4	5
Easy to learn	3	4	5	4	4
Usability	5	4	5	5	4
Support	5	5	3	2	3
Extensibility	5	5	3	3	5
Reliability	5	5	5	5	5
Installation	5	5	5	4	5
Data Preprocessing	5	3	3	3	5
Data Visualization	4	3	5	5	3
Total Score	41	39	37	35	39

Table 2.4: Comparison of Data Mining tools in terms of system features

N.B: The system feature evaluation was based on a 5-point scale, with higher scores indicating better results such as high/comprehensive/easy/simple, and lower scores for negative results such as low/none/difficult/complex. Accordingly RapidMiner scored the maximum than the rest.

As expected, each of the five open source data mining tools evaluated has its unique set of pros and cons. Concerning the data mining functionalities Weka and Rapid Miner performed well in comparison with the rest of the tools and for the case of user friendliness all of the selected tools have a graphical user interface to communicate with the users so that all the tools do not have problem over this parameter. But most significantly the parameters placed under system features have a leading role in identifying the better mining tool. In sight of this criterion RapidMiner and WEKA are better than the rest under consideration.

Taking in to consideration the results seen in the above comparisons, WEKA and RapidMiner mining tools have shown a better performance than the rest. But due to the investigator's prior knowledge, popularity on previously done health related researches and due to having a well documented manual, WEKA has been selected to run the experiments [9, 48, 53, 54].

2.1.7. Healthcare and Data Mining

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models [29]. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns [30].

Data mining is not a new concept, rather it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling.

In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Several factors have motivated the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders [31]. Fraud detection using data mining applications is prevalent in the commercial world, for example, in the detection of fraudulent credit card transactions. Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection [30].

Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data [32]. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care [33]. Healthcare organizations that perform data mining are better positioned to meet their long-term needs, as Benko and Wilson argue [34]. Data can be a great asset to healthcare organizations, but they have to be first transformed into information.

Yet another factor motivating the use of data mining applications in healthcare is the realization that data mining can generate information that is very useful to all parties involved in the healthcare industry. For example, data mining applications can help healthcare insurers detect fraud and abuse, and healthcare providers can gain assistance in making decisions, for example,

in customer relationship management. Data mining applications also can benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices [35, 36].

There are also other factors boosting data mining's popularity. For instance, as a result of the Balanced Budget Act of 1997, the Centers for Medicare and Medicaid Services must implement a prospective payment system based on classifying patients into case-mix groups, using empirical evidence that resource use within each case-mix group is relatively constant. CMS has used data mining to develop a prospective payment system for inpatient rehabilitation [37].

2.1.8. HIV/AIDS and ART

2.1.8.1. HIV/AIDS

The human immunodeficiency virus, or HIV, is the virus that causes HIV infection. During HIV infection, the virus attacks and destroys the infection-fighting CD4 cells of the body's immune system. Loss of CD4 cells makes it difficult for the immune system to fight infections. Acquired immunodeficiency syndrome, or AIDS, is the most advanced stage of HIV infection [13].

2.1.8.2. CD4 Cells Count

HIV attacks the immune system, destroying the system's infection-fighting CD4 cells. Keeping the immune system healthy is an important goal of HIV treatment.

The CD4 count measures the number of CD4 cells in a sample of blood. The CD4 count of a healthy person ranges between 500 and 1,200 cells/mm3. An HIV-infected person with a CD4 cells count of less than 200 cells/mm3 develops AIDS [39].

Because a falling CD4 count is a sign that HIV is damaging the immune system, the test is used to monitor HIV infection. Once treatment is started, the CD4 count is also used to monitor the effectiveness of anti-HIV medications [39].

Once the patient starts treatment, he/she should have a CD4 count once every 3 to 4 months. An increasing CD4 count is a sign that the immune system is recovering. If patient's regimen is working well, he/she need a CD4 count only once every 6 to 12 months [39].

2.1.8.3. Role of ART on HIV

In the early 1980s when the AIDS epidemic began, people living with HIV were not likely to live more than a few years. However, with the development of safe and effective drugs, HIV positive people now have longer and healthier lives [40].

Currently available drugs do not cure HIV infection but they do prevent the development of AIDS [40]. They can stop the virus being made in the body and this stops the virus from damaging the immune system, but these drugs cannot eliminate HIV from the body. Hence, people with HIV need to continuously take antiretroviral drugs.

The use of ART in combinations of three or more drugs as HIV treatment has dramatically improved the quality of life for people with HIV and prevented them from dying early, since 1996 in countries where they are widely accessible [40].

2.1.8.4. WHO Clinical Stages

The WHO clinical stages are made to support the delivery of medications to adults and adolescents having HIV. The following details about the clinical stages of HIV are based on the WHO guidelines revised for 2010 [41].

Clinical stage 1

Asymptomatic

Persistent generalized lymphadenopathy

Clinical stage 2

Moderate unexplained weight loss (under 10% of presumed or measured body weight)

Recurrent respiratory tract infections (sinusitis, tonsillitis, otitis media, pharyngitis)

Herpes zoster

Angular cheilitis

Recurrent oral ulcerations

Popular pruritic eruptions

Seborrhoeic dermatitis

Fungal nail infections

Clinical stage 3

Unexplained severe weight loss (over 10% of presumed or measured body weight)

Unexplained chronic diarrhoea for longer than 1 month

Unexplained persistent fever (intermittent or constant for longer than 1 month)

Persistent oral candidiasis

Oral hairy leukoplakia

Pulmonary tuberculosis

Severe bacterial infections (e.g. pneumonia, empyema, meningitis, pyomyositis, bone or joint infection, bacteraemia, severe pelvic inflammatory disease)

Acute necrotizing ulcerative stomatitis, gingivitis or periodontitis

Unexplained anaemia (below 8 g/dl), neutropenia (below 0.5 x 109/l) and/or chronic thrombocytopenia (below 50 x 109/l)

Clinical stage 4

HIV wasting syndrome

Pneumocystis jiroveci pneumonia

Recurrent severe bacterial pneumonia

Chronic herpes simplex infection (orolabial, genital or anorectal of more than 1 month's duration or visceral at any site)

Oesophageal candidiasis (or candidiasis of trachea, bronchi or lungs)

Extrapulmonary tuberculosis

Kaposi sarcoma

Cytomegalovirus disease (retinitis or infection of other organs, excluding liver, spleen and lymph nodes)

Central nervous system toxoplasmosis

HIV encephalopathy

Extrapulmonary cryptococcosis including meningitis

Disseminated nontuberculous mycobacteria infection

Progressive multifocal leukoencephalopathy

Chronic cryptosporidiosis

Chronic isosporiasis

Disseminated mycosis (histoplasmosis, coccidiomycosis)

Recurrent septicaemia (including nontyphoidal *Salmonella*)

Lymphoma (cerebral or B cell non-Hodgkin)

Invasive cervical carcinoma

Atypical disseminated leishmaniasis

Symptomatic HIV-associated nephropathy or HIV-associated cardiomyopathy

2.1.8.5. WHO Guidelines on Initiating ART

Antiretroviral drug treatment guidelines have changed over time [42]. Prior to 1987, no antiretroviral drugs were available and treatment consisted of treating complications from the immunodeficiency [42]. After antiretroviral medications were introduced, most clinicians agreed that HIV positive patients with low CD4 counts should be treated, but no consensus formed as to whether to treat patients with high CD4 counts [42].

In 1995, David Ho promoted a "hit hard, hit early" approach with aggressive treatment with multiple antiretrovirals early in the course of the infection [43]. Later reviews noted that this approach of "hit hard, hit early" ran significant risks of increasing side effects and development of multidrug resistance, and this approach was largely abandoned [44].

The timing of when to initiate therapy has continued to be a core controversy within the medical community [45]. The development of a stable consensus is hampered by the lack of randomized controlled studies with many guidelines and consensus statements basing their recommendations on observational studies [45]. More recently, the trend has been in favor of earlier treatment of

asymptomatic HIV patients, with more studies analyzing various treatment regimens in progress [45].

There is a consensus among experts that, once initiated, antiretroviral therapy should never be stopped [46]. This is because the selection pressure of incomplete suppression of viral replication in the presence of drug therapy causes the more drug sensitive strains to be selectively inhibited. This allows the drug resistant strains to become dominant [46]. This in turn makes it harder to treat the infected individual as well as anyone else they infect.

WHO recommends that HIV infected adolescents and adults should start antiretroviral therapy when the following conditions are met [47]:

1. It is recommended to treat all patients with CD4 counts of ≤350 cells/mm3 irrespective of the WHO clinical stage.
   (Strong recommendation, moderate quality of evidence)
2. It is recommended that all patients with WHO clinical stage 1 and 2 should have access to CD4 testing to decide when to initiate treatment.
   (Strong recommendation, low quality of evidence)
3. It is recommended to treat all patients with WHO clinical stage 3 and 4 irrespective of CD4 count.
   (Strong recommendation, low quality of evidence)

2.2. Review of Related Works

2.2.1. Application of Data Mining on HIV/AIDS Datasets

A study done by Elias [48] indicated that data mining techniques can be used to develop HIV status predictive model. In that study, J48 and ID3 decision tree implementation classification algorithms were used. And also Apriori association rule generation algorithm was used to show the existence of association between the dependent variable HIV status and the independent variables.

The study has used 51270 records and 17 selected attributes to train and test the models. Moreover, the study is organized in to five experimental scenarios for both J48 and ID3 classifiers to build up the respective models. In addition to the classification mining tasks, association rule discovery was made to uncover the underlying association between the predicted variable HIV status and the predicting variables.

The findings of Elias [48] study have shown that, the classification rules revealed that females are more vulnerable to HIV than Males. The other two classification rules were generated in terms of age group of clients. According to the study, age group 25-49 were the most susceptible subset of the population and age group 50 and above were also becoming vulnerable to HIV/AIDS as the patterns have indicated. In addition to the above mentioned, clients whose previous HIV status was negative has also shown negative during repeated testing.

Group of the association rules indicated that there is a direct relationship between never married clients and HIV negative status. This is often applicable, if their previously HIV status is negative. The other observed association rule is those clients whose primary reason to visit the center labeled risk is associated with positive HIV status. Moreover the study revealed that the attributes age, economic status, and knowledge about the disease are among the contributing factors to engage in risk which may amount to HIV positive result.

Another study conducted by Vararuk et al [49] showed the use of data mining technology to investigate patterns in HIV/AIDS patient data, so that these patterns can be used for better management of the disease and more appropriate targeting of resources.

The study has incorporated a total of 250,000 anonymised records from HIV/AIDS patients in Thailand and then imported into a database. IBM's Intelligent Miner was used for clustering and association rule discovery.

The findings of this study stipulated that clustering technique has highlighted groups of patients with common characteristics and also the errors found within the data. Association rules identified associations that were not expected in the data and were different from traditional reporting mechanisms utilized by medical practitioners. It has also allowed the identification of symptoms that co-exist or are precursors of other symptoms.

A study conducted by Rosma M. et al [50] has indicated that it is very possible to predict aids survival of patients by using data mining techniques. In that study a predictive data mining techniques based on fuzzy regression concept was used for the prediction. Data from 997 patients diagnosed with HIV/AIDS and treated at the University Hospital, Kuala Lumpur, Malaysia from 1987 to 2007 were used in the study. Out of 997 patients, 831 patients were male while 166 patients were female. However, after data cleansing and preprocessing, only 298 cases were used. The data were divided into training and testing sets. 248 cases (83%) were used for training purposes and the remaining 50 cases (17%) were used for testing. The performance of fuzzy regression technique to model the survival of AIDS was compared with that of fuzzy neural network. Both models demonstrated high accuracy prediction. Percentage accuracy of FuReA's prediction on AIDS survival shows that CD4, CD8 and viral load counts could be the appropriate predictors/markers to be used for predicting AIDS survival due to the high accuracies demonstrated.

In conclusion the researcher explained that Fuzzy regression prediction technique had been shown to be competent and had produced outstanding results. The experiment on predicting AIDS survival demonstrated the feasibility of applying FuReA, a fuzzy regression data mining technique on some significant medical cases.

2.2.2. Application of Data Mining on ART Datasets

The very first work which triggered the researcher's internal motive to attempt this research is the unpublished research done by Behailu G [9]. The study was done on constructing CD4 status predictive model. In his research the following attributes were used from the database: RegistrationDate, Sex, Age, ReligionID, MaritalStatusID, EducationalLevelID, Occupation,

ARTStatus, Functional Status, ELDate, EligibleReasonID, ARTStartDate, FamilyPlanning, PregnantYN, OAWeight, OAWHO stage, CurrentRegimen, PastARVTreatment, and OACD4 (Baseline CD4 count). In his finding, he tried to show the level of prediction of these attributes in terms of predicting the dependent variable OACD4 count to classify a patient to either "Low" CD4 status or "Normal" CD4 status.

According to this study, among the attributes ranked for predicting CD4 status of patients following ART, Eligible reason is the first determining attribute and the least determining attribute is Educational level. The ten top determining attributes are Eligible reason, ART status, ART start year, OA weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status and Past ARV consecutively.

The findings of this study indicated that, the best performance is achieved by the J48 decision tree algorithm with a setting of generalized decision tree with pruning and reduced attributes. The model classified 88.79% instances correctly and remaining 11.21% incorrectly. The weighted average precision of the model is 0.88 with recall of 0.89 and ROC area of 0.85. The generated tree contains 760 leaves with tree size of 916. The analysis of this model shows that the model is quit efficient to predict CD4 status of patients following ART.

Another study done in South Africa by Yashik et al [10], which is entitled "Support Vector machines to forecast changes in CD4 count of HIV-1 positive patients" showed that it is possible to mathematically forecast a change in CD4 count using SVM. The best accuracy achieved was 83% using genome, current CD4 count and number of weeks from baseline CD4 count. The remaining 17% of misclassified CD4 count changes are distributed such that the majority of the discrepancies are one category/CD4 change group apart.

This pilot study forms part of a larger study to create a web-based HIV resistance portal. It is envisioned that this portal will be used to guide treatment of complicated HIV resistant patients by prompting clinicians to enter genomic, virological and treatment history data and then providing them with information about the specific patient's current resistance profile, future resistance profiles, the effect of changes in treatment and the prediction of the onset of AIDS, opportunistic diseases and mortality. Forecasting the CD4 count of a patient from genotypic information is thus vital to the creation of the resistance portal, which will guide the clinician in determining the optimal therapy for individual patients.

A study conducted by Zazzi M et al [71] on prediction of response to antiretroviral therapy by human experts and the EuResist data-driven expert system. The EuResist system was compared with 10 HIV-1 drug resistance experts for the ability to predict 8-week response to 25 treatment cases derived from the EuResist database validation data set.

All current and past patient data were made available to simulate clinical practice. The experts were asked to provide a qualitative and quantitative estimate of the probability of treatment success. Out of the given 15 treatment successes and 10 treatment failures, the number of mislabeled case were 6 for the Euresist and 6 to 13 for that of human experts [mean $\pm$ standard deviation i.e. 9.1 $\pm$ 1.9]. The accuracy of EuResist is higher than the average of human experts (0.76 vs. 0.64, respectively). The quantitative estimates computed by EuResist were significantly correlated (Pearson r = 0.695, $P < 0.0001$) with the mean quantitative estimates provided by experts. This indicates that, the EuResist engine performed better than the human experts.

In a study conducted by Michael Lee et al [51], a novel method for predicting maximum recommended therapeutic dose (MRTD) is presented using quantitative structure property relationships (QSPRs) and ANNs. In the study MRTD data of 31 structurally diverse ARVs were collected from FDA MRTD Database or package inserts. Molecular property descriptors of each compound, that is, molecular mass, aqueous solubility, lipophilicity, biotransformation half life, oxidation half life, and biodegradation probability were calculated from their SMILES codes. A training set (n = 23) was used to construct multiple linear regression and back propagation neural network models. The models were validated using an external test set (n = 8) which demonstrated that MRTD values may be predicted with reasonable accuracy. Model predictability was described by root mean squared errors (RMSEs), Kendall's correlation coefficients (tau), P-values, and Bland Altman plots for method comparisons. MRTD was predicted by a 6-3-1 neural network model (RMSE = 13.67, tau = 0.643, P = 0.035) more accurately than by the multiple linear regression (RMSE = 27.27, tau = 0.714, P = 0.019) model. Both models illustrated a moderate correlation between aqueous solubility of antiretroviral drugs and maximum therapeutic dose. MRTD prediction may assist in the design of safer, more effective treatments for HIV infection.

CHAPTER THREE
DATA MINING ALGORITHMS AND RESEARCH METHODOLOGY

3.1. Mining Algorithms Used in the Study

As it has been discussed in the literature review, **Section 2.1.5**, this study has incorporated classification mining algorithms to build the model. Accordingly J48 from decision tree, PART from rule induction, SMO from support vector machine and MLP from artificial neural network are selected to run the experiments due to their popularity in recent publications [9, 10, 53].

3.1.1. Decision Tree

A decision tree is a classifier expressed as a recursive partition of the instance space [55]. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called a "root" that has no incoming edges [55]. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an "internal" or "test" node [55]. All other nodes are called "leaves" (also known as "terminal" or "decision" nodes). In the decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values [55]. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector (affinity vector) indicating the probability of the target attribute having a certain value [55].

3.1.1.1. J48

The basic algorithms for decision tree induction is a greedy algorithm which constructs decision trees in a top down approach dividing each node recursively until a leaf node is encountered [14]. The following algorithm shows the generation of a decision tree from a training tuples of data partition [14].

Input:

- *Data partition, D, which is a set of training tuples and their associated class labels;*
- *Attribute list, the set of candidate attributes;*
- *Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.*

Output: **A decision tree**.

Method:

Create a node N;
If tuples in D are all the same class, C then
Return N as a leaf labeled with the class C;
If attribute list is empty then Return N as a leaf node labeled with the majority class in D;
Apply attribute selection method (D, attribute list) to find the "best" splitting criterion;
Label node N with splitting criterion;
If splitting attribute is discrete valued and multi way splits allowed then
Create list ← attribute list splitting attribute; // remove splitting attribute
For each outcome j of splitting criterion
Let D_j be the set of data tuples in D satisfying outcome j;
If D_j is empty then Attach a leaf labeled with the majority class in D to node N;
Else attach the node returned by Generate decision tree (D_j, attribute list) to node N;
End for
Return N

To construct optimal decision tree, Entropy and Information Gain needs to be calculated. The information gain measure enables to select the test attribute at each node in the tree and the attribute with the highest information gain or greatest entropy reduction is chosen as the test attribute for the current node [25].

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (C_1, C_2, C_3 . . . , C_m). Let Si be the number of sample of S in class C_i. The expected information needed to classify a given sample is given by:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(P_i) \quad \ldots\ldots\ldots\ldots (3.1)$$

Where P_i is the probability that an arbitrary sample belongs to class C_i and a log function is to the base 2 is used because the information is encoded in bits. The entropy, or expected information based on the partitioning into subsets by A is given by:

$$D(n_+, n_-) = -\frac{n_+}{n}\log_2\frac{n_+}{n} - \frac{n_-}{n}\log_2\frac{n_-}{n} \quad \ldots\ldots\ldots\ldots\ldots (3.2)$$

The smaller the entropy value is, the greater the purity of the subset partitions [3]. The information that would be gained by branching on A is given by the following formula:

$$\text{Gain}(A) = I(s_1, s_2, s_3 \ldots, s_m) - \text{Entropy}(A) \quad \ldots\ldots\ldots\ldots\ldots (3.3)$$

This algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly [25]. Decision tree use the above formulas to determine which attribute to split on. The highest information gain considered as the test attribute for the given database [25]. Then, a node is created and branches out. The same procedure followed for the next attribute to split.

3.1.2. Rule Induction

There is an alternative approach to rule induction that avoids global optimization but nevertheless produces accurate, compact rule sets [14]. The method combines the divide-and-conquer strategy for decision tree learning with the separate-and-conquer one for rule learning. It adopts the separate-and-conquer strategy in that it builds a rule, removes the instances it covers, and continues creating rules recursively for the remaining instances until none are left [14].

3.1.2.1. PART

PART (Partial Decision Tree) is a rule induction algorithm which grabs rule from a decision tree. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees [14]. To generate such a tree, the construction and pruning operations are integrated in order to find a "stable" sub tree that can be simplified no further [14]. Once this sub tree has been found,

tree building ceases and a single rule is read off. The following algorithm depicts the steps and procedures followed in implementing PART rule induction.

Initialize **E** to the instance set
For each class **C**, from smallest to largest
BUILD:
Split E into Growing and Pruning sets in the ratio 2:1
Repeat until (a) there are no more uncovered examples of C; or (b) the description length (DL) of ruleset and examples is 64 bits greater than the smallest DL found so far, or (c) the error rate exceeds 50%:
GROW phase: Grow a rule by greedily adding conditions until the rule is 100% accurate by testing every possible value of each attribute and selecting the condition with greatest information gain G
PRUNE phase: Prune conditions in last-to-first order. Continue as long as the worth W of the rule increases
OPTIMIZE:
GENERATE VARIANTS:
For each rule **R** for class **C**,
Split E afresh into Growing and Pruning sets
Remove all instances from the Pruning set that are covered by other rules for C
Use **GROW** and **PRUNE** to generate and prune two competing rules from the newly split data:
R1 is a new rule, rebuilt from scratch;
R2 is generated by greedily adding antecedents to R.
Prune using the metric A (instead of W) on this reduced data
SELECT REPRESENTATIVE:
Replace R by whichever of R, R1 and R2 has the smallest DL.
MOP UP:
If there are residual uncovered instances of class C, return to the BUILD stage to generate more rules based on these
CLEAN UP:
Calculate DL for the whole rule set and for the rule set with each rule in turn omitted; delete any rule that increases the DL
Remove instances covered by the rules just generated
Continue

3.1.3. Support Vector Machine

In machine learning, **support vector machines** (**SVMs**, also **support vector networks**) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [56]. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [56]. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other [56]. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [56].

In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [56].

Since 1992 Support Vector Machine (SVM) were largely unnoticed due to widespread belief in the statistical and/or machine learning community, despite being theoretically appealing [57]. A due attention towards SVM has come in to prominence when excellent results achieved in numeral recognition, text categorization and computer vision; today SVM show better results than neural network and other statistical models [57].

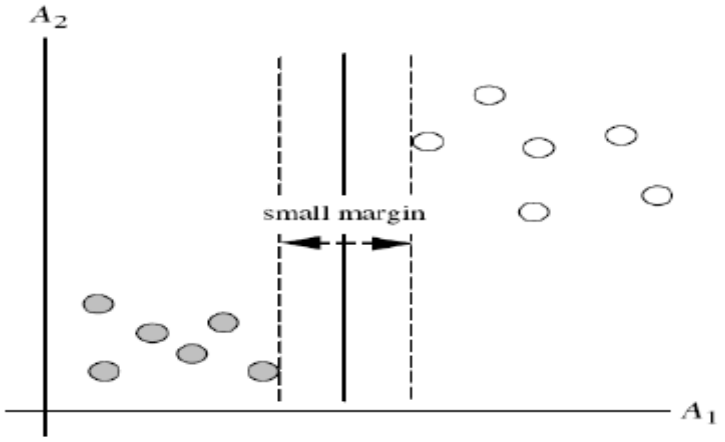

Figure 3.1: Linearly Separable 2D Training Data

To explain SVM; if it is a two-class problem where the classes are linearly separable, an algorithm is implemented to finds a special kind of linear models. Let $\mathbf{x_i \in R_n}$, **(i= 1, 2, 3… m)** represents the vectors and $\mathbf{y_i \in \{1,-1\}}$. The term $f(x_i)$ can be represented by a linear functions of the form by $y_i = f(x_i)$.

$$f(x_i) = (W.\ X) + b \quad \text{.....} \quad (3.4)$$

Where W is a weight vector namely, W= {W_1, W_2, W_3 . . . W_n} and b is a scalar, often referred to as bias [57]. There is infinite number of hyper planes/separating lines that could be drawn for classifying the two-classes [57]. To find the optimal linear model or Hyper plane (n dimensions) that will have the minimum classification error on previously unseen tuples, SVM search for maximum marginal hyper plane [57].

3.1.3.1. SMO (Sequential Minimal Optimization)

SMO implements John Platt's sequential minimal optimization algorithm for training a support vector classifier [14]. This implementation globally replaces all missing values and transforms nominal attributes into binary ones [14]. It also normalizes all attributes by **default**. In that case the coefficients in the output are based on the normalized data, not the original data, this is important for interpreting the classifier [14].

3.1.4. Artificial Neural Network

An artificial neural network (ANN) tries to capture the brain problem solving ability and apply them to information systems. The human brain provides proof of the existence of massive neural networks that can succeed at those cognitive, perceptual, and control tasks in which humans are successful [58]. Humans are better than computers in performing complex tasks such as speech and image recognition, but computers are faster in performing mathematical computations [25]. ANN attempts to replicate the human massive processing capacity to problem solving information systems [25].

3.1.4.1. Multilayer Feed-Forward Neural Network

The back propagation algorithm performs learning on a *multilayer feed-forward* neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-

forward neural network consists of an *input layer*, one or more *hidden layers*, and an *output layer* [25]. An example of a multilayer feed-forward network is shown in **Figure 3.2**.

As clearly noted by Han et al [25], each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuron like" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples [25]. The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer. It is fully connected in that each unit provides input to each unit in the next forward layer [25].

The steps for computing the output of a single neuron are as follows:

1. Compute the weighted sum of inputs to the neuron and add the bias to the sum

$$(I_j) = \sum_i W_{ij} O_i + \theta_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.5)$$

   Where W_{ij} is the weight of the connection from unit i in the previous layer to unit j; O_i is the output of unit i from the previous layer, θ_j is the bias of unit which acts as a threshold.

2. Each unit in the hidden and output layers takes its net input and then applies an activation function [25]. The output of the activation function is defined to be the output of the neuron.

$$O_j = \frac{1}{1 + e^{-Ij}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.6)$$

   This function is called logistic or sigmoid function or also is referred to as a squashing function, because it maps a large input domain onto the smaller range of 0 to 1 [25].

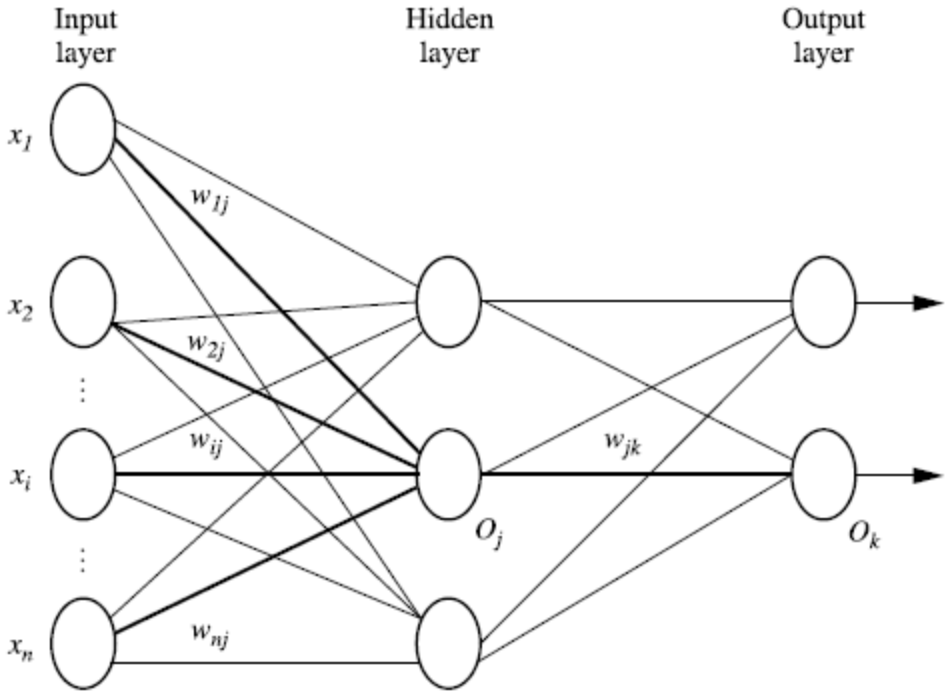

Figure 3.2: A Multilayer Feed Forward Neural Network

3.1.5. Ensemble Learning

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem [64]. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypothesis and combine them to use [64]. The well known machine learning and data mining based classification algorithms use probabilistic methods, rule-based learners, linear models such as neural networks and support vector machines, decision trees and instance-based learners. Further, a combination of different classification algorithms can result in improved classification accuracy [65]. The commonly used ensemble techniques are ***bagging***, ***boosting***, ***voting*** and ***stacking***. Typically, an ensemble is constructed in two steps. First, a number of base learners are produced, which can be generated in a ***parallel*** style or in a ***sequential*** style where the generation of a base learner has influence on the generation of subsequent learners. Then, the base learners are combined to use, where among the most popular combination schemes are ***majority voting*** for classification and ***weighted averaging*** **for regression** [64].

For this study, a boosting technique is selected to boost the classifier's weak performance in each of the experiments. AdaBoostM1 is an extension of AdaBoost algorithm which is available under the Weka package and thus it has been selected to boost the experiments done by using the above base classifiers.

3.2. Model Evaluation Parameters

Data mining problems involving classification, it is very common to measure classifiers performance in terms of the error rate or misclassification rate [14]. The classifier predicts class label of each instances and if it is correct, it is counted as success, else counted as error. Evaluating the accuracy using training datasets derive a classifier or predictor to be likely misleading due to overspecialization of the learning algorithms to the data [14, 25]. For this reason it is better to assess the error rate based on independent test dataset that have no role in classifier datasets. Both, training data and the test data, needs to be representative sample of the problem [14].

For measuring accuracy of a classifier, there are a number of techniques such as the holdout, random sub-sampling, bootstrap and k-fold cross-validation, where the dataset is divided in to training and testing to train and test the classifier respectively [25].

The holdout method reserves a certain amount of instances for training and uses the remainder for testing (and sets part of that aside for validation, if required) [14]. In practical terms, it is common to hold out one-third of the data for testing and use the remaining two-thirds for training.

Bootstrap is based on the statistical procedure of sampling with replacement [14]. The bootstrap procedure may be the best way of estimating the error rate for very small datasets.

In K-fold cross validation technique, one decide the number of fold (partitions of the data) and then the data is split in to K approximately equal partitions; and thus K-1/K and 1/K partition in turn used for training and testing the classifier respectively [14]. 10 fold cross validation is the most commonly used data partitioning technique for training and testing a classifier. Extensive tests on numerous different datasets, with different learning techniques, have shown that ten is the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up [14]. Although these arguments are by no means conclusive, and

debate continues to rage in machine learning and data mining circles about what is the best scheme for evaluation, tenfold cross-validation has become the standard method in practical terms [14]. Accordingly, ten-fold cross validation is selected for this research to train and test the classifier models.

3.2.1. Confusion Matrix

Confusion matrix is a tool for analyzing how well the classifier can recognize tuples of different classes [14]. Give m classes, a confusion matrix is a table of at least m by m. For instance, in tow class case with classes yes and no, sick or healthy, cancer or not cancer, or lend or not lend, a single prediction has four different possible outcomes.

		Predicted Class	
		Yes	**No**
Actual Class	**Yes**	**TP**	**FN**
	No	**FP**	**TN**

Table 3.1: Different Outcomes of a Two-Class Prediction

In the above table four possible outcomes of a two class classifier are observed, where true positive (TP) and true negative (TN) are the correct classifications. A false positive (FP) is when the outcome is incorrectly predicted as *yes* (or positive) when it is actually no (negative). A false negative (FN) is when the outcome is incorrectly predicted as negative when it is actually positive [14, 25]. Confusion matrix enables the mechanism to understand how well the classifier has classified the tuples as "yes" and "no".

On the other hand sensitivity (the true positive rate) and false positive rate can also be computed using this confusion matrix to rate the performance of a classifier [25]. The equations placed below (**Eq3.7** and **Eq3.8**) can be used to compute the sensitivity and false positive rate of a classifier respectively.

$$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \quad \text{.....} \quad (3.7)$$

$$\textbf{FPR} = \frac{\textbf{FP}}{\textbf{TN + FP}} \quad \text{……..…… (3.8)}$$

The other parameter used in this study to compare the performance of classifiers is “precision”; which indicates the percentage of instances classified as positives by the learned model and that are actually positives. The equation presented below can be used to compute the precision of a given classifier.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}} \quad \text{…… (3.9)}$$

The **F-value** metric is one measure that combines the trade-offs of ***precision*** and ***recall,*** and outputs a single number reflecting the "goodness" of a classifier in the presence of rare classes [73]. The F-value represents the trade-off among different values of **TP,** FP, and **FN**. The expression for the F-value is as follows:

$$\text{F-Value} = \frac{(1+\beta^2)*recall*precision}{\beta^2*precicion+recall} \quad \text{……..…..… (3.10)}$$

Where β corresponds to the relative importance of ***precision*** versus ***recall.*** It is usually set to 1.

The entire success of the classifier or accuracy of the classifier is the number of correctly classified instances divided by the total number of instances. On the other hand the error can be computed through subtracting the accuracy from one [14]. The overall accuracy of the classifier can be computed by using the equation below (**Eq3.9**).

$$\textbf{Over all Accuracy} = \frac{\textbf{TP + TN}}{\textbf{TP + TN + FP + FN}} \quad \text{……. (3.11)}$$

In this research the performance of each model is weighted in terms of sensitivity, FPR, F-measure, mean absolute error, AUC and accuracy values. Based on the remark from this comparison, the one with better performance was discussed in brief and associated prototype has been developed.

3.2.2. ROC Curve

ROC curves are a useful visual tool for comparing two classification models. The acronym stands for **receiver operating characteristic**, a term used in signal detection to characterize the tradeoff between hit rate and false-alarm rate over a noisy channel [25]. ROC curves depict the performance of a classifier without regard to class distribution or error costs [25]. They plot the true positive rate on the vertical axis against the false positive rate on the horizontal axis [25]. The former is the number of positives included in the sample, expressed as a percentage of the total number of positives (**TP Rate = 100 × TP/ (TP + FN)**); the latter is the number of false positives included in the sample, expressed as a percentage of the total number of negatives (**FP Rate = 100 × FP/ (FP + TN)**). A sample ROC curve representing the percentage of true positives and false positives is presented in the figure below. The plot also shows a diagonal line where for every true positive of such a model, there is more likely to encounter a false positive. Thus, the closer the ROC curve of a model is to the diagonal line, the less accurate the model. If the model is really good, initially we are more likely to encounter true positives as we move down the ranked list. Thus, the curve would move steeply up from zero. Later, as we start to encounter fewer and fewer true positives, and more and more false positives, the curve cases off and becomes more horizontal.

To assess the accuracy of a model, we can measure the area under the curve. The closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy will have an area of 1.0.

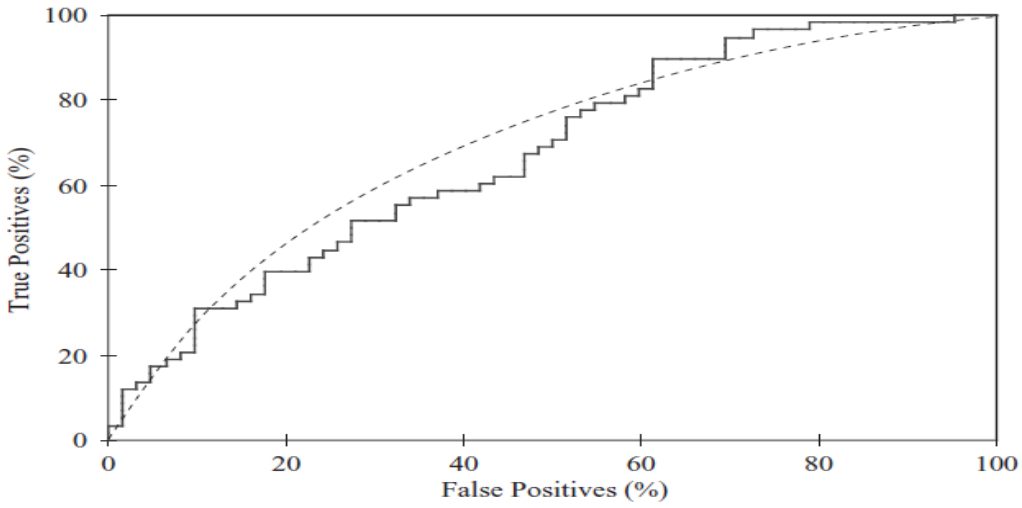

Figure 3.3: Sample ROC curve

3.3. Research Methodology

3.3.1. Area of the study

The study is conducted in three government owned hospitals namely Jimma, Bonga and Aman which are found in the south west part of Ethiopia at a distance of 349Km, 449Km and 584Km from Addis Ababa respectively. Bonga and Aman hospitals are administered under the southern Nations and Nationalities Peoples Regional state where as Jimma hospital is governed under Oromia regional state. This area is selected due to the fact that large number of the younger people who are much more vulnerable to HIV [63] lives around by engaging themselves in nearby government and private work environments; like 'Wush Wush Tea Plantation' which is around Bonga, 'Bedele Beer factory' which is near Jimma, 'Gold mining sites in Dima' which is around Aman, 'Bebeka Coffee Plantation' which is again not far from Aman and other private and government owned investment sites around the region.

3.3.2. Dataset size

In this study patient records of ART following adults (those who are above 18 years of age) with attributes which are assumed to be significant in determining the CD4 count of patients over a period of time were selected from the ART dataset of those three hospitals. Accordingly from Aman hospital among a total of **5,600** registered patient's records in the ART clinic, **3,600** of them which were under treatment were taken. From Jimma hospital dataset among **6,750** registered ART following patient records **3,000** of them were selected. From Bonga dataset, **652** records were selected among **2,500** registered cases. Joining the selected records from those three hospitals, a total of **7,252** records were obtained in the study to train and test the classifier model.

3.3.3. Research Design (Hybrid-DM)

As per the discussion made on model assessment in literature review **section 2.1.3**, Hybrid-DM process model was selected as a better model regarding its design to suit for academic researches. Accordingly, to realize a model that yields optimum classifier of CD4 status of an individual, a Hybrid data mining process model was used to guide the overall execution of the project. Hybrid-DM model is developed by Cios et al [24] based on the CRISP-DM model by

adopting it to academic research. As Cios et al [24] urges, the main differences and extensions with CRISP-DM are:

- It provides 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.

Moreover hybrid model is selected due to its recent works done in the health domain [53, 66] and that has showed wonderful results. This has played a leading role in the selection of this model besides its appropriateness for academic researches.

The Hybrid-DM model presented in **Figure 3.4** below consists of a six-step knowledge discovery process which includes; understanding the problem domain, understanding data, preparation of data, data mining, evaluation of the discovered knowledge, use of the discovered knowledge [24].

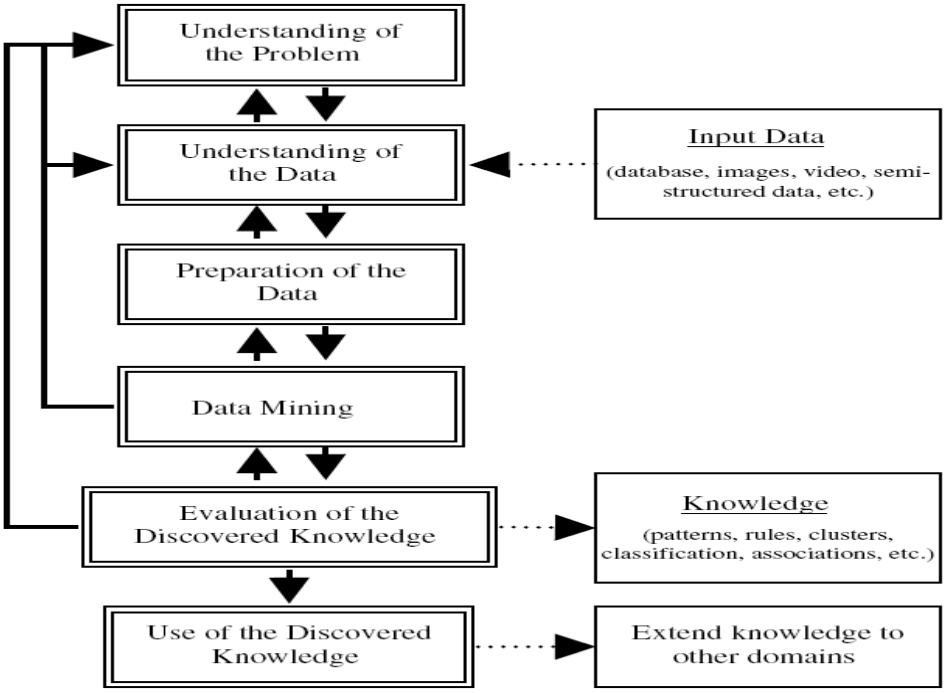

Figure 3.4: Hybrid-DM Process Model

3.3.3.1. Problem understanding

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.

3.3.3.2. Data understanding

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.

3.3.3.3. Data Preparation

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. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in **Step 1**.

3.3.3.4. Data Mining

As it has been discussed in **section 3.1 above,** the investigator has employed three classification data mining algorithms to develop and compare the classification models. Accordingly, among the available algorithms in WEKA machine learning software; J48, PART, MLP, and SMO were used in this research.

Since there are three dependent variables to be predicted in this study and also classification algorithms can only handle one dependent variable at a time, therefore the mining tasks were performed step wise for each outcome variable. This is to mean that, the first outcome variable gets the first sight to get done and then the turn goes to the second and finally to the third outcome variable. While running experiments for the second outcome variable, the first outcome

variable (CD4 count after six months) is used as a predicting variable. For the experiments done to predict the third outcome variable (CD4 count after eighteen months), the previous two outcome variables (CD4 count after six months and CD4 count after twelve months) are used together with the ten predicting variables. This is done due to the justification that, an individual who is tested to know his/her sixth month CD4 count receives a counseling service in order to sustain the better counts scored or to improve the poor counts. This has its own influence in the later CD4 counts. The integrated rules are made by joining the best rules selected from each of the three best models of the three outcome variables.

3.3.3.5. 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 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. Accordingly the discovered patterns were checked for their interestingness together with the domain area experts and also the interpretations were critically commented by experts.

3.3.3.6. 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 has been documented.

The results of this thesis work will be disseminated to the following stakeholders and to any interested parties.

- It will be presented to the schools of information science and public health of Addis Ababa University.
- A hardcopy of the documentation will be available in the bibliographic library of the school information science.
- Maximum effort will be exerted to publish the result on different journals to initiate other interested groups to find gaps and do more research in the area.

CHAPTER FOUR

DATA PREPARATION

Data Mining is a technology that uses various techniques to discover hidden knowledge from a data stored in large databases, data warehouses and other massive information repositories [25]. To discover non-trivial knowledge and patterns, the database must undergo effective data preparation to bring a valid output.

The data mining model used in this study is a Hybrid-DM, which is a six step knowledge discovery process. Among these phases, the first three, business understanding, data understanding and data preprocessing are meant to prepare the data for data mining tasks. As noted in Han et al [25], data understanding and preprocessing usually consumes the majority of the effort in the entire data mining process.

4.1. Business Understanding

Business understanding is a stage where the researcher is acquainted with the overall business process or working conditions of the identified work process. To this end, the researcher has made efforts to understand what the business is and its in and out as much as possible. This study was done in consultation with domain area experts to have an in-depth insight into the problem domain and also physical observation of the situation was done to have a real picture of the problem. The domain area experts constitute physicians engaged in counting CD4 cells (T-lymphocyte cells), nurses who are engaged in giving counseling services and taking care to those admitted to inpatient department, people working under Federal Ministry of Health on policy development, monitoring and evaluation of its implementation, and ART data clerks.

Business understanding is mainly concerned with the determination of business objectives, assessment of the situations, and determination of data mining goals.

4.1.1. Identifying Business Objectives

Antiretroviral therapy is a therapy given to suppress the human immunity virus load in HIV positive patients [6]. With this regard the primary objective of the ART clinics is to check how much the patient's body has responded to the ARV drug during therapy. The other objective of these clinics is to ensure that the given drug regimen is appropriate to the given individual based

on the changes seen in the patient's CD4 cells count. To achieve these two objectives the hospitals are currently using a machine to count the patient's CD4 in six month intervals.

Generally, the overall goal of ART is to improve the quality of life of HIV positive individuals through continuous provision of the antiretroviral drugs and in the mean time counting their CD4 cells within the interval of six months so as to affirm their immunological advancement over time.

4.1.2. Determination of Data Mining Goals

The data mining goals are derived from the business objectives so that the mining tasks were done targeting the stated goals. Accordingly the following data mining goals were identified to guide the upcoming events like experimentation and prototype development. **Goal 1**: Given the demographic data, baseline WHO clinical stage and baseline Cd4 cells count, predict the CD4 cells count of a patient at six, twelve and eighteen months of therapy. **Goal 2**: From the identified predicting variables, determine those having a better prediction performance.

4.2. Data Understanding

Data is a core element in data mining projects. Therefore, for a data mining project to succeed data understanding is a mandatory step to bring about a valid result. In order to meet the general objective of this research, collecting representative subset of ART data is a prerequisite. Data understanding begins with collection of the initial dataset. The data collection was done in three hospitals namely, Jimma, Bonga and Aman which are located in the south western part of Ethiopia.

Productive discussions were made with domain area experts working in FMOH and physicians working on ART program within the mentioned hospitals in order to select the final dataset to be used for the study. In addition to domain area experts, literatures done on related areas were analyzed to authenticate the selected attributes. Accordingly ten (10) independent variables i.e. Age, Sex, Marital Status, Educational Status, Family Planning usage status, Pregnancy status, Functional status, Baseline WHO stage, Drug regimen and Baseline CD4 count and three (3) outcome variables i.e. CD4 count after six months of treatment, CD4 count after twelve months treatment and CD4 count after eighteen months of treatment were selected from the huge dataset

containing above eighty five (85) attributes from each of the three datasets. The detail of selected attributes is presented in the following table.

No	Attribute's Name	Description	Values	Data Types
1	Age	Age of a patient in years	Numeric age values	Numeric
2	Sex	Sex of a patient	Male, Female	Nominal
3	Marital Status	Marital Status of a patient	Never married, Married, Separated, Divorced, Widow	Nominal
4	Educational Status	Educational status of a patient	No education, Primary, Secondary, Tertiary	Nominal
5	Family Planning	Family planning usage status of a patient	Yes, No	Nominal
6	Pregnancy Status	Pregnancy status of a patient	Yes, No, Not Applicable (NA)	Nominal
7	Functional Status	Functional Status of a patient	W-working, A-Ambulatory, B-Bedridden	Ordinal
8	Baseline WHO Stage	WHO clinical stage of the disease when the patient starts the drug	One, Two, Three, Four	Ordinal
9	Drug Regimen	The drug regimen given at the very onset of treatment	1-a = d4T - 3TC - NVP 1-b = d4T - 3TC - EFV 1-c = AZT - 3TC - NVP 1-d = AZT - 3TC - EFV 1-e = TDF - 3TC - NVP 1-f = TDF - 3TC - EFV 1-g = d4T - 3TC - ABC 2-a = ABC – ddl – Lpv/r 2-b = ABC – ddl – NFV 2-c = TDF – ddl – Lpv/R 2-d = TDF – ddl – NFV	Nominal
10	Baseline CD4 Count	The CD4 count taken when the patient begins the drug	Numeric values ranging from 0 to 935	Numeric
11	CD4 Count After Six Months	CD4 count taken after six months of therapy	Numeric values ranging from 4 to 992	Numeric
12	CD4 Count After Twelve Months	CD4 count taken after twelve months of therapy	Numeric values ranging from 17 to 1108	Numeric
13	CD4 Count After Eighteen Months	CD4 count taken after eighteen months of therapy	Numeric values ranging from 15 to 1083	Numeric

Table 4.1: Selected Attributes with their Description

4.3. Data Preprocessing

Today's real-world databases are highly susceptible to noisy, missing, incomplete and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources [25]. High quality data will lead to high-quality mining results and vice verse [25]. Consequently, real world data of low quality needs preprocessing.

Different data preprocessing tasks were involved in this study including data cleaning, handling outliers', data integration, data transformation, and data reduction techniques. Data integration is another major task, which merges data from multiple sources into a single and coherent data sources. Data transformation, such as normalization and discretization are also used to optimize the accuracy and efficiency of mining algorithms.

Data reduction is also another important task which is meant to reduce the data size in order to obtain quick processing time and save a memory space to be consumed.

The overall objective of data preprocessing is to obtain a data prepared in the form required by the data mining algorithms and to expose as much information as possible for data modeling.

4.3.1. Data Integration

Initially, a total of **10,220** tuples were collected from the three hospitals' ART databases based on their status to be enrolled on ART or those who are started taking ARV. Accordingly, **7,252** records are only selected to develop the classifiers by leaving others which are not complete enough to address the intended objective. The data collected from Bonga and Aman Hospitals have similar structure as they are developed using Microsoft Access. These two hospitals are supported by a non-governmental organization called JHU-TSEHAI both technically and materially. But in the case of Jimma Hospital, ART service is supported by ICAP-Ethiopia whereby the database is totally different from that of Bonga and Aman ART databases which uses Microsoft SQL server to store and interface with the users. Therefore, a great deal of effort was exerted and continuous discussions were made with domain experts and people working on HIV/AIDS prevention and control in order to integrate the three datasets and come up with a uniform dataset capable of achieving the study objectives.

The data integration was made by first selecting the most determinant attributes found in both of the two different databases and then removing the remaining irrelevant attributes as suggested by experts. After completing this task, the three data sets were joined together using Microsoft Excel version 2007.

4.3.2. Exploratory Data Analysis

Descriptive data summarization techniques can be used to identify the typical properties of your data and highlight which data values should be treated as noise or outliers [25]. Furthermore, missing values can easily be identified through this technique which in turn facilitates the next phases of data preparation.

To understand the nature of the data values in the selected ART dataset an exploratory data analysis was done both for numeric and categorical attributes. Descriptive statistics comprising of the total number of valid and missing instances, minimum, maximum, mean and standard deviation were computed to show the distribution of values for the numeric attributes. Categorical attributes were explored to expose the valid and missing instances and the percentage distribution of each data instances.

In this study, five (5) numeric and eight (8) categorical variables including the dependent variables were used to reach at the final goal of this study. The following table presents numeric variables and their distribution in the dataset.

Exploratory Data Analysis							
No.	**Attribute**	**Valid**	**Missing**	**Min**	**Max**	**Mean**	**Standard Deviation**
1	Age	7252	0	18	81	32.51	8.571
2	Baseline CD4 Count	7252	0	0	935	155.72	110.226
3	CD4 Count After Six Months	7252	0	4	992	281.05	145.017
4	CD4 Count After Twelve Months	7252	0	17	1108	327.2	147.339
5	CD4 Count After Eighteen Months	7252	0	15	1083	364.16	153.472

Table 4.2: Descriptive Statistics of Numeric Attributes

The remaining eight attributes which are ordinal and nominal types are presented in the following frequency distribution table. The frequency of data values in terms of number and percentage together with the number of valid and missing values is presented in the table below.

No.	Attribute	Data Values										Valid	Missing
1	Sex	**Male**				**Female**						7252 (100%)	No Missing
		3030 (41.8%)				4222 (58.2%)							
2	Marital Status	**Divorced**	**Married**	**Never Married**	**Separated**	**Widow**						6658 (91.8%)	594 (8.2%)
		574 (7.9%)	3470 (47.8%)	1400 (19.3%)	452 (6.2%)	762 (10.5%)							
3	Educational Status	**No Education**	**Primary**	**Secondary**	**Tertiary**							6654 (91.8%)	598 (8.2%)
		1192 (16.4%)	2348 (32.4%)	2426 (33.5%)	688 (9.5%)								
4	Family Planning	**Yes**			**No**							6652 (91.7%)	600 (8.3%)
		4066 (56.1%)			2586 (35.7%)								
5	Pregnancy Status	**Yes**			**No**	**NA**						7252 (100%)	No Missing
		148 (2%)			4074 (56.2%)	3030 (41.8%)							
6	Functional Status	**Ambulatory**	**Bed Ridden**	**Working**								6784 (93.5%)	468 (6.5%)
		2176 (30%)	288 (4%)	4320 (59.6)									
7	Baseline WHO Stage	**One**	**Two**	**Three**	**Four**							7084 (97.7%)	168 (2.3%)
		1148 (15.8%)	1560 (21.5%)	3468 (47.8%)	908 (12.5%)								
8	Drug Regimen	1-a	1-b	1-c	1-d	1-e	1-f	1-g	2-a	2-c	2-d	7252 (100%)	No Missing
		3,070	768	1,101	520	1420	298	60	1	12	2		

Table 4.3: Frequency Distribution of Nominal Attributes

Therefore, looking at this frequency distribution table, it is very clear to define the characteristics of patients' in terms of these variables. For example, if we take the variable "Family Planning ", among the study subjects taken from the source dataset, 56.1% of them use family planning mechanisms to protect the virus transmission to others or formation of mutation of the viral strains where as the remaining 35.7% do not use such prevention technique.

4.3.3. Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent [59]. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data [59]. In this study the above mentioned data cleansing mechanisms were used to prepare a clean dataset that can easily be operated by the data mining algorithms.

4.3.3.1. Handling Inconsistent Data

Following the completion of data integration the next operation performed was identifying the completeness of the dataset. Considering the follow up data, only those having a complete follow up data, i.e. those having CD4 count taken at six, twelve and eighteen months of therapy are selected to involve in the model building. Consequently, from initially collected 10220 records, 7252 tuples were selected as valid records for further data preprocessing.

The other inconsistency was observed in the "WHO stage" field where the valid data values are one (1), two (2), three (3) and four (4). But some tuples were represented by a value which is inconsistent with the standards set to represent WHO stage. Thus, those data values were automatically deleted.

4.3.3.2. Handling Outliers

Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data [60]. The outlying values may arise due to typographic or measurement error like the value of a nominal attributes is misspelled or in a numeric attributes case correct ones can be possibly filled with incorrect values which results in outliers that can be easily figured out by graphing one variable at a time [60]. Further cause to noisy data can be faulty data collection instruments, data entry problems, data transmission problems, and inconsistencies in naming convention are the typical ones for noisy data to happen [60].

Outliers can be detected using visualization techniques (i.e. using histograms for numerical and bar charts for categorical attributes) so that the values which are far away from the normal values in case of numeric attributes and unusual data value in the case of categorical attributes are depicted as outlier. According to Chakrabarti et al [59], a common rule of thumb for identifying suspected outliers is to single out values falling at least **1.5 * IQR** above the third quartile or

below the first quartile i.e. if it is lower than **Q1 – 1.5(IQR)** or higher than **Q3 + 1.5(IQR)**, where **IQR** = **Q3 – Q1**.

No	Attribute Name	Min	Q1	Median	Q3	Max	IQR	Outliers
1	Age	18	26	30	37	81	11	< 9.5 and > 54
2	Baseline CD4 Count	0	74	139	220	935	146	Above 439 ~ 440
3	CD4 Count After Six Months	4	183	255	348	992	165	Above 596 ~ 600
4	CD4 Count After Twelve Months	17	228	302	394	1108	166	Above 598 ~ 600
5	CD4 Count After Eighteen Months	15	260	352	436	1108	176	Above 700

Table 4.4: Statistical Summary of Numeric Attributes

From the above table, it is straight forward to pinpoint the values under each attributes lying in the region of outliers and those which are not. Consequently, the cut off points identified in each of these attributes were used in defining an **outlier class** for the independent variable "**Age**". The variable "**Age**" is already trimmed for its lower values by limiting only those above 18 years of age to enroll in the study, so that the remaining outlier values found above **Q3 + 1.5*IQR** (i.e. **54**) were treated as one group for the experiments. But here the grouping was made based on the WHO standard so that the age range above **50** was taken as one group due to less number of instances above **54** which is marked as outlier age group. Moreover, the mean value is calculated by ignoring outlier values in order to avoid the impact on it.

According to Chakrabarti et al [59] Box plots are popular ways of visualizing a distribution in numeric type of data values. A box plot incorporates the five-number summary as follows:

- Typically, the ends of the box are at the quartiles, so that the box length is the inter quartile range (IQR = Q3 - Q1).
- The median is marked by a line within the box.
- Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Accordingly this study has used box plots to visualize outliers found within those five (5) numerical attributes. But for the case of categorical attributes by looking at the frequency

distribution presented above (**Table 4.3**), one can easily understand that, the categorical variables are free from outliers.

The following box plot depicts the distribution of the numeric data values within the experimentation dataset.

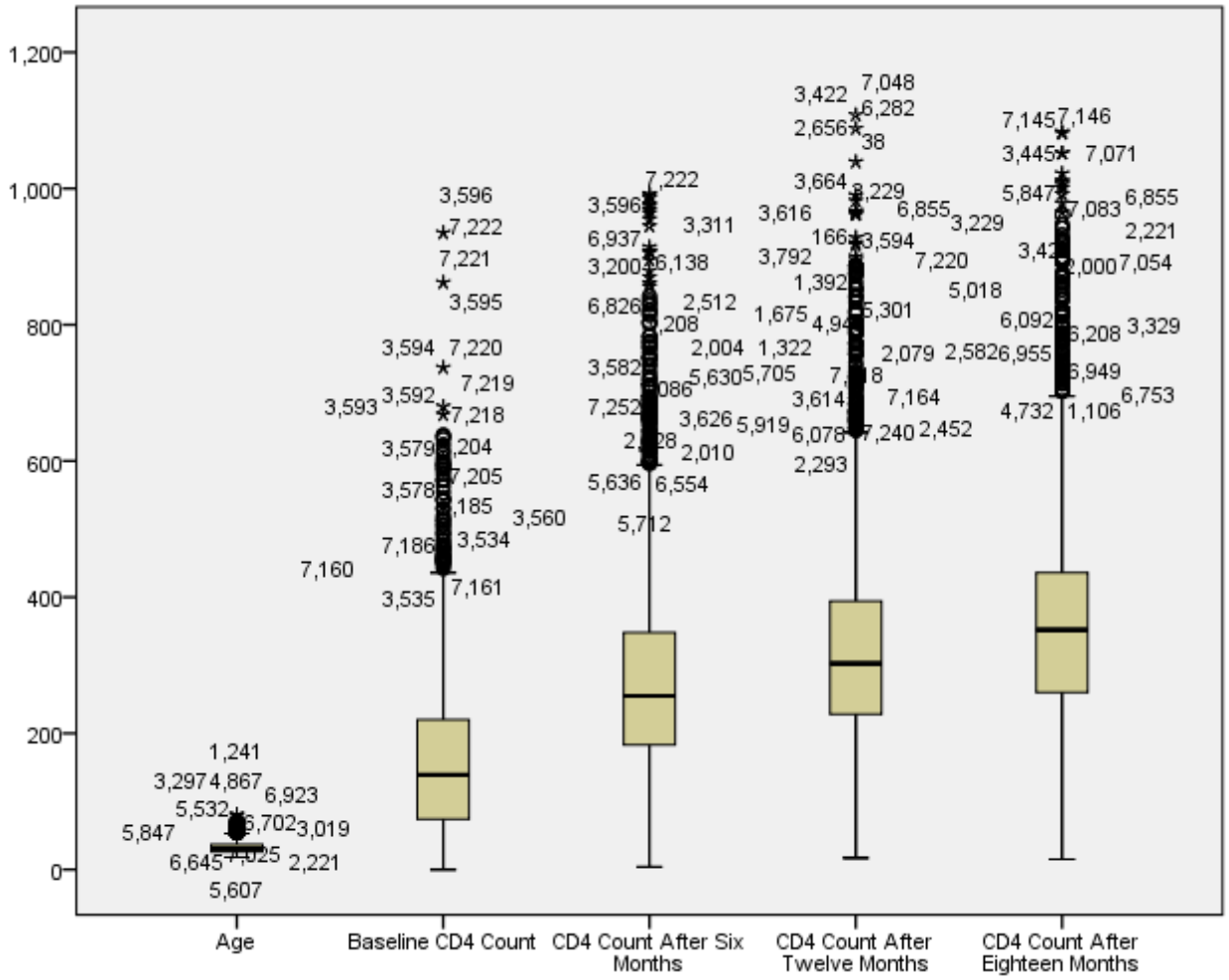

Figure 4.1: Box Plot of the Dataset to Visualize Outliers

4.3.3.3. Handling Missing Values

Missing data is a problem that continues to affect data analysis methods [60]. Even as our analysis methods gain sophistication, we continue to encounter missing values in fields, especially in databases with a large number of fields [60]. The absence of information is rarely beneficial. All things being equal, more data is almost always better.

According to T. Larose [60] Missing values may occur for several reasons. Such as mal-functioning measurement equipment, lack of consistency with other recorded data and thus deleted, or respondents in a survey may refuse to answer certain questions such as age or income and data may not be recorded due to misunderstanding. But those missing values needs to be given significant attention.

To deal with missing values, alternatives are suggested by T. Larose [60] and Chakrabarti et al [59]. These are:

- Ignore the missing value
- Replace the missing value manually
- Replace the missing value with a global constant to fill in the missing value
- Replace the missing value with some constant, specified by the analyst
- Replace the missing value with the field mean(for numerical variables) or the mode (for categorical variables)
- Replace the missing values with a value generated at random from the variable distribution observed.

In the study dataset, missing values were observed in both the numeric and categorical variables as it has been presented in **Table 4.2** and **Table 4.3** above. Fortunately, only six (6) attributes (which are Marital Status, Educational Status, Family Planning, Functional Status, Baseline WHO Stage and original regimen) has missing value from the categorical variables with no missing from the numerical attributes. Hence, the percentage of missing values in the mentioned attributes is not considerably significant, thus the missing values were replaced by the mean for numeric attributes using SPSS package version 20 and the mode for categorical attributes using Microsoft Excel version 2007 before applying data mining algorithms. The following table (**Table 4.5**) depicts attributes with the number of missing values, mean, mode and the action taken to replace the missing values.

No	Attribute Name	Valid	Missing	Mean	Mode	Action Taken
1	Marital Status	6658 (91.8%)	594 (8.2%)		Married	Replaced by the mode
2	Educational Status	6654 (91.8%)	598 (8.2%)		Secondary	Replaced by the mode
3	Family Planning	6652 (91.7%)	600 (8.3%)		Yes	Replaced by the mode
4	Functional Status	6784 (93.5%)	468 (6.5%)		Working	Replaced by the mode
5	Baseline WHO Stage	7084 (97.7%)	168 (2.3%)		Stage-3	Replaced by the mode
6	Drug Regimen	7229 (99.7%)	23 (0.3%)		1-a = D4T + 3TC + NVP	Replaced by the mode

Table 4.5: Handling Missing Values

4.3.4. Data Reduction

According to Chakrabarti et al [59], 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. Strategies for data reduction includes data cube aggregation, attribute subset selection, dimensionality reduction, numerosity reduction, discretization, and concept hierarchy generation.

In this study, among the listed data reduction techniques attribute subset selection, discretization, and dimensionality reduction were used to prepare a data that can increase efficiency of the mining operation.

4.3.4.1. Discretization

As noted in Chakrabarti et al [59], data discretization techniques are used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data [59]. This leads to a concise, easy-to-use, knowledge-level representation of mining results. Among the available data discretization techniques', binning has been selected to discretize the numeric attributes in to ranges of values for the independent variables and in to classes for the three dependent variables.

Binning is a top-down splitting technique based on a specified number of bins [59]. The attribute values can be discretized by applying equal-width or equal-frequency binning. The investigator in collaboration with domain area experts and expert from the ministry of health approved standard based binning to avoid the possible deviation from the already accepted standards.

Accordingly, "Age" is binned in to seven groups with an equal bin width of five with difference only made to the initial group (18-24) and the final outlier group (Above 50). The rest data values between 25 and 49 are equally partitioned in to five groups to form a total of seven bins.

Baseline CD4 count is classified according to the WHO cutoff point (200) to initiate ART in resource limited settings. With this regard, the continuous valued baseline CD4 counts are categorized in to two as "Below 200" and "Above 200" for the values less than 200 and greater than or equal to 200 respectively. But due to the large number of instances categorized under the category below 200 (68%), instances under this category are further classified in to three as "Below 50", "50 - 99" and "100 – 199". This categorization is done based on a recommendation of experts from FHAPCO. Moreover, a national study conducted by FHAPCO [72] indicated that more than 80% of ART following patients initiated therapy below 200.

But for the three outcome variables, categorization is made by following a WHO guideline prepared on case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children [61]. The following table presents the immunological classification of HIV patients.

HIV Associated Immunodeficiency	**Age related CD4 values**			
	<11 Months (%CD4+)	**12 – 35 Months (%CD4+)**	**36 – 59 Months (%CD4+)**	**>5 years (Absolute number per mm3 or %CD4+)**
None or not significant	>35	>30	>25	**>500**
Mild	30 – 35	25 – 30	20 – 25	**350 – 499**
Advanced	25 – 29	20 – 24	15 – 19	**200 – 349**
Severe	<25	<20	<15	**<200 or <15%**

Table 4.6: WHO Immunological Classification for Established HIV Infection

The detail of discretization employed on each of the continuous variables is shown in **Table 3.6** below.

No.	Attribute's Name	Min	Max	Outlier's cutoff point	Bin Size	Resulting Range of Values
1	Age	18	81	54	7	18 – 24, 25 – 29, 30 – 34, 35 – 39, 40 – 44, 45 – 49, and Above 50
2	Baseline CD4 Count	0	935	440	4	<50, 50 – 99, 100 - 199 and >=200
3	CD4 Count After Six Months	4	992	600	4	<200, 200 – 349, 350 – 499, >=500
4	CD4 Count After Twelve Months	17	1108	600	4	<200, 200 – 349, 350 – 499, >=500
5	CD4 Count After Eighteen Months	15	1108	700	4	<200, 200 – 349, 350 – 499, >=500

Table 4.7: Discretization of Continuous Numeric Attributes

4.3.4.2. Concept Hierarchy Generation

Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior) [25]. Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret. This contributes to a consistent representation of data mining results among multiple mining tasks, which is a common requirement. In addition, mining on a reduced data set requires fewer input/output operations and is more efficient than mining on a larger, ungeneralized data set [25]. In this study the attribute "Original Regimen" contains ten unique instances which are complex enough to the data mining algorithms and still difficult to interpret the final result. Accordingly, due to the less number of second line regimen users in the dataset, all regimens found in this category were combined together to form one generalized group called "**Second-Line**". Therefore, all the regimen types labeled 2-a, 2-b and 2-c has been changed with the general form of the group "**Second-Line**". This has reduced the number of instances from ten to eight.

4.3.4.3. Dimensionality Reduction

In dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data [59]. If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called **lossless** [59]. Here, the investigator has used data encoding to transform the data represented in characters to digits, so that the original data is represented with a reduced form without losing any detail of the original data.

The data encoding done on the study variables is depicted in the table below.

No	Attribute's Name	Old Value	New Value
1	Age	{18 – 24, 25 – 29, 30 – 34, 35 – 39, 40 – 44, 45 – 49, and Above 50}	{1, 2, 3, 4, 5, 6, 7}
2	Sex	{Female, Male}	{1, 2}
3	Marital Status	{Never Married, Married, Separated, Divorced, Widow}	{1, 2, 3, 4, 5}
4	Educational Status	{No Education, Primary, Secondary, Tertiary}	{1, 2, 3, 4}
5	Family Planning	{NO, YES}	{0, 1}
6	Pregnancy Status	{NO, YES}	{0, 1}
7	Functional Status	{B, A, W}	{1, 2, 3}
8	Baseline WHO Stage	{One, Two, Three, Four}	{1, 2, 3, 4}
9	Drug Regimen	{1-a, 1-b, 1-c, 1-d, 1-e, 1-f, 1-g, Second-Line}	{1, 2, 3, 4, 5, 6, 7, 8}
10	Baseline CD4 Count	{<50, 50 – 99, 100 – 199, >=200}	{1, 2, 3, 4}
11	CD4 Count After Six Months	{<200, 200 – 349, 350 – 499, >=500}	{1, 2, 3, 4}
12	CD4 Count After Twelve Months	{<200, 200 – 349, 350 – 499, >=500}	{1, 2, 3, 4}
13	CD4 Count After Eighteen Months	{<200, 200 – 349, 350 – 499, >=500}	{1, 2, 3, 4}

Table 4.8: Data Encoding of Continuous Numeric Attributes

4.3.4.4. Attribute Subset Selection

The major criterion for selecting an attribute set at this initial stage is to check whether each attribute is relevant to the data mining objective. Two crows corporation [62] suggests that usefulness to the data mining objective is the major criteria in selecting attributes at the initial

stage. Therefore, based on literatures and necessary information obtained from domain experts the researcher has identified ten socio-demographic and clinical assessment attributes which can potentially predict CD4 count.

The attribute's rank is computed using Weka's ChiSquaredAttributeEval attribute evaluation algorithm, so that the importance of all attributes is ordered based on their chi-square statistics. The ability of chi-square to deal with categorical variables makes it the choice of this study because the selected attributes are all nominal valued. Accordingly, all attributes in the three experimental setups were selected to enroll in all of the experiments due to having a higher chi square value which is an indication of the attribute's association with the dependent variable. The selected attributes for the three outcome variables along with their information gain value are appended in *Appendix A, B and C*.

4.3.5. Data Transformation

According to Han et al [25], data transformation involves transforming or consolidating the data to a form appropriate for mining. Data transformation usually involves data smoothing, generalization of data, normalization of data, aggregation of data, and attribute construction.

In this study, among the above listed data transformation techniques data normalization is used in order to normalize the data values to fall between 0 and 1 for ANN experiments. Therefore, for experiments done using ANN the data values are normalized in between 0 and 1 to make the algorithm efficient and effective.

CHAPTER FIVE

DATA MINING AND MODEL SELECTION

Data Mining is the fourth stage in hybrid-dm process model which is discussed in depth in the third chapter, **Section 3.3.3**. In this study, four data mining algorithms were used to achieve the objective of developing predictive model using patient records taken from ART dataset of Jimma, Bonga and Aman Hospitals. Decision tree (J48), rule induction (PART), artificial neural network (Multilayer Perceptron) and support vector machine (Simple Minimal Optimization) were used to run the experiments. In addition to these four algorithms, a boosting algorithm called AdaBoostM1 was used to improve the minimum accuracy scored by those base classifiers. In this section of the study three major experimentations were done for each of the three outcome variables by using the above mentioned algorithms. Ten fold cross validation technique is also used to train and test classifiers in each experimental scenario. Different model selection schemes were used to evaluate model's performance. The detail of these tasks is presented in the next sections.

5.1. Experimental Setup

All experiments are done on the final dataset which has passed all the preprocessing operations to suit the selected mining algorithms. The dataset contains **7,252** records with **ten** predicting and three outcome variables. Initially, the dataset was in spreadsheet format and then converted in to CSV in order to be read by WEKA machine learning software. To ease the repeated access of the file, it has been converted in to ARFF using Weka. Moreover, these experiments are done by using the current stable version of Weka, which is version 3.7.7. The final dataset with its selected attributes is shown in Figure 5.1 below.

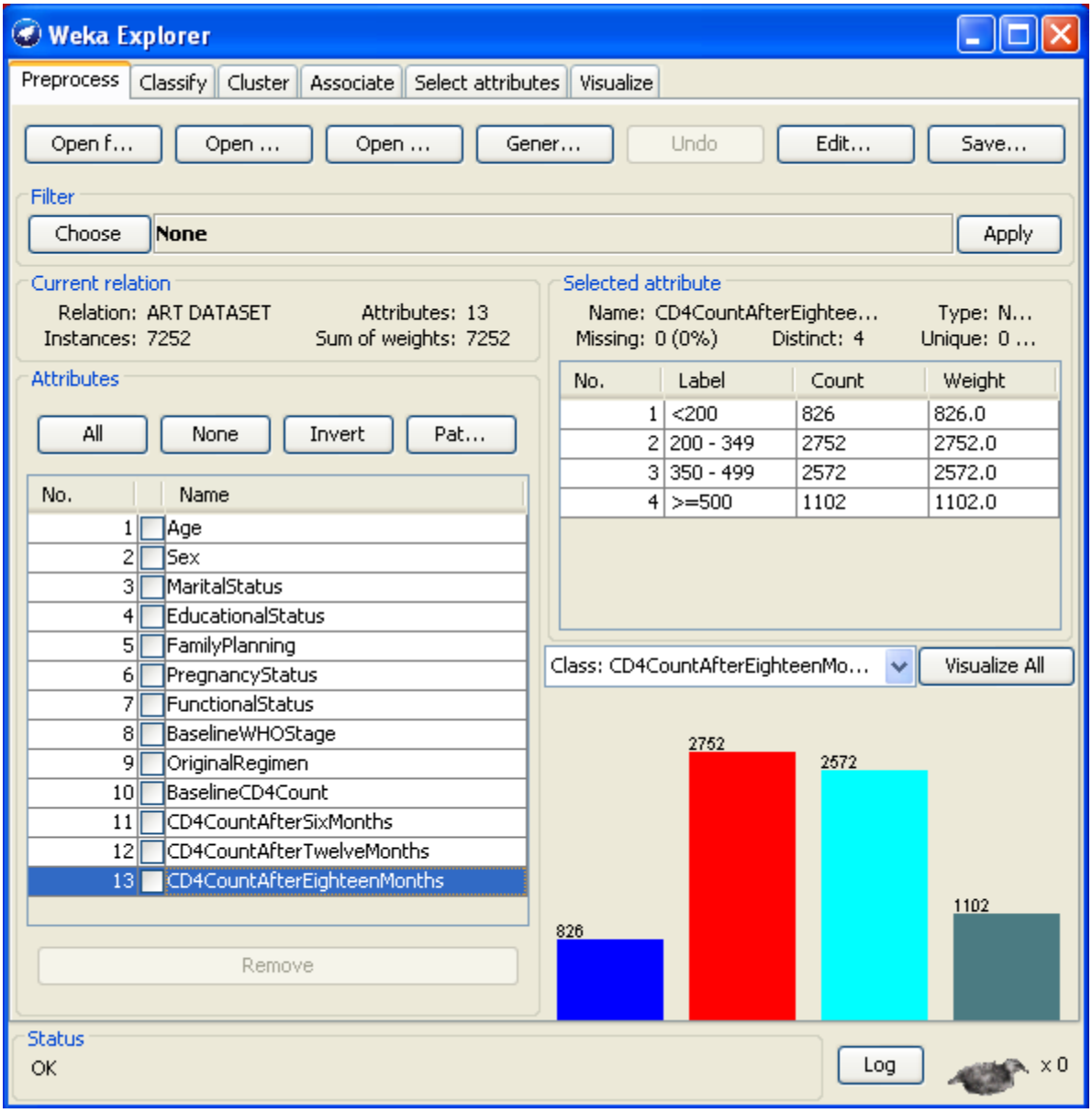

Figure 5.1: WEKA View of the Final Dataset with the Selected Thirteen Attributes

In this study, the outcome variables are three and each of them contains four similar classes with which the predictions are made. The big challenge here is each of the three outcome variables contains imbalanced classes which can possibly deteriorate the classifier's predictive accuracy.

Recent reports from both academy and industry indicate that, the imbalanced class distribution of a data set poses serious difficulty to most classifier learning algorithms which assume a relatively balanced distribution [67, 68]. Imbalanced class distribution is characterized as that there are many more instances of some classes than others. With imbalanced data, classification rules that

predict the small classes tend to be fewer and weaker than those that predict the prevalent classes; consequently, test samples belonging to the small classes are misclassified more often than those belonging to the prevalent classes.

Sun Y et al [69] argue that classification of data with imbalanced class distribution creates a significant drawback on the performance to be attained by most standard classification learning algorithms which assumes a relatively balanced class distribution and equal misclassification costs. But there are developments in machine learning to address issues associated to developing classifier models for multiclass problems. Sun Y et al [69] indicated that, multiclass boosting algorithms namely AdaBoostM1 (Adaptive Boosting Algorithm for Multiple classes, which is an extension of AdaBoost) and AdaC2M1, which is an extension of AdaBoostM1 performed better than the base classifier C4.5 algorithm. These two algorithms together with the base classifier were tested by using three different datasets taken from UCI information repository to test the predictive performance and thus the finding of this report indicated that AdaC2M1 is a little bit better than AdaBoostM1 and these two are by far better than the base classifier algorithm.

Accordingly, due to the availability of the Meta classifier AdaBoostM1 algorithm under the Weka package, the researcher has used this algorithm together with the four base classifier algorithms to run experiments.

The class distribution of each outcome variable is presented in the table below.

No.	Dependent Variable	Class	Number of Instances	% Distribution
1	Sixth Month CD4 Count	**<200**	2268	31.27%
		200 – 349	3188	43.96%
		350 – 499	**1218**	**16.8%**
		>=500	**578**	**7.97%**
2	Twelfth Month CD4 Count	**<200**	**1246**	**17.18%**
		200 – 349	3364	46.39%
		350 – 499	1834	25.29%
		>=500	**808**	**11.14%**
3	Eighteenth Month CD4 Count	**<200**	**826**	**11.39%**
		200 – 349	2752	37.95%
		350 – 499	2572	35.46%
		>=500	**1102**	**15.2%**

Table 5.1: Class Distribution of the Three Outcome Variables

The experiments conducted in this study are categorized in to three major categories as experimentation to model sixth month CD4 count, experimentation to model twelfth month CD4 count and experimentation to model eighteenth month CD4 count. Under these three blocks of experimental designs each of the four mining algorithms with different settings and the boosting algorithm with its default parameters are used to conduct experiments in order to obtain one model with a better accuracy in each case.

The above mentioned experimentations with their design are presented in the following sections with brief explanations on the setup and observed findings.

5.2. Experimentations to Model Sixth Month CD4 Count

To come up with one selected model, lots of experiments were done using all of the four mentioned algorithms by changing the parameters contained in each of them. Accordingly, in those experiments done for the first outcome variable (CD4 count after six months), all of the ten (10) predicting variables are used. The experimental setup and details of the findings for each of the selected mining algorithm is presented in the following section.

5.2.1. J48 Experiments

Two experiments are conducted using J48 by switching the parameter with pruning to TRUE and FALSE to form two separate experimental settings. The Meta classifier algorithm (AdaBoostM1) is used in both scenarios to evaluate the performance gained in those minority classes. This indicates that a total of four experiments were conducted in this category. The confidence factor in all scenarios is made to be **0.5**, which is found to be a better value after attempting successive experiments at different confidence levels. It is also confirmed by different researchers for its better accuracy than taking the default confidence value **0.25** [9, 66].

Setting #1: J48 Experiment with All Attributes and with pruning

Setting #2: J48 Experiment with All Attributes and without pruning

Setting #3: AdaBoostM1 Experiment with its default parameters and taking J48 with pruning as a base classifier

Setting #4: AdaBoostM1 Experiment with its default parameters and taking J48 without pruning as a base classifier

In the first scenario of this experiment, the **10** attributes and **7,252** records are used by taking the default parameter value with pruning. The result showed that, the experiment has generated a model with a tree size of **4,136** and **3,096** leaves.

In the second scenario, the same number of attributes and records are used to run the experiment. But relatively larger tree having a size of **4,492** and **3,365** leaves is generated.

In the third and fourth experimental setups a classifier with relatively better predictive accuracy is generated due to the addition of the boosting algorithm. The tree and leaf size of the third and fourth scenarios decreased to **1,613** and **1,270**, and **2,102** and **1,638** respectively.

The performance of the four experiments is presented in the table below.

Experiment	**Accuracy**	**Tree Size**	**Leaf Size**	**WTP Rate**	**WFP Rate**	**WPrecision**	**WF-measure**	**WROC Area**
J48 pruned	62.73%	4,136	3,096	0.627	0.192	0.622	0.623	0.873
J48 unprunned	65.80%	4,492	3,365	0.658	0.172	0.656	0.656	0.896
AdaBoostM1 with pruned J48	81.67%	1,613	1,270	0.817	0.093	0.816	0.816	0.941
AdaBoostM1 with unprunned J48	**82.10%**	2,102	1,638	**0.821**	**0.09**	**0.821**	**0.821**	**0.945**

Table 5.2: J48 Experiments Performance Evaluation for the Sixth month CD4 Count

As indicated in the table above, the accuracy of all models indicate the classifier's ability of classifying new instances and it is calculated to be: 62.73% with misclassification of 37.27% for the first scenario; 65.8% of correct classification with misclassification of 34.2% for the second scenario; 81.67% of correct classification with error rate of 18.23% for the third scenario and the final scenario has revealed 82.1% correct classification and error rate of 17.9%.

The false positive rate indicated in each model shows the percentage of records which are wrongly classified in to any of the four classes. Accordingly, the last model has wrongly classified 9% of the records and hence it is the least in this category.

The ROC area also indicates the area under the axis of true positive and false positive rates. Therefore, as the area under the ROC curve gets larger, it indicates that the classifier is putting

more true positives than false positives in the given class. The boosted unprunned J48 decision tree has scored a better performance taking the performance parameters indicated above. Therefore, the **boosted and unprunned J48** decision tree has been selected to be compared with other classifiers generated under this category.

In a table presented under *Appendix D*, the *TPR*, *F-measure* and *area under the ROC curve* of the minority classes (**350 – 499** and **>=500**) has dramatically increased whereas the FPR has lowered to a smaller value which indicates that less number of instances are wrongly classified under those minority and majority classes.

5.2.2. PART Experiments

Like the J48 experiments done above, PART experiments are also done in four experimental settings based on the parameter pruning and the boosting algorithm. Here also the confidence factor is made to be 0.5 for its better accuracy results. The experimental settings are indicated according to the next scenario.

Setting #1: PART Experiment with pruning.

Setting #2: PART Experiment without pruning.

Setting #3: AdaBoostM1 Experiment with its default parameters and taking PART with pruning as a base classifier.

Setting #4: AdaBoostM1 Experiment with its default parameters and taking PART without pruning as a base classifier.

In the first setting of PART experiment a classifier with an accuracy of 58.41%, weighted TPR of 58.4% and weighted FPR of 20.8% is generated by taking the default value of pruning. In addition to these performance parameters, the model has generated a total of **1,477 rules** to represent the patterns found within the dataset.

In the second setting, the same number of attributes and records are used by switching the default parameter value of unprunned to "TRUE". This experiment has produced a classifier with an accuracy of 61.56% and also it has relatively higher TPR and lower FPR than the first experiment.

In the third experiment, a boosting algorithm is applied on the base classifier with the pruning state turned on. Accordingly, a classifier with better accuracy (81.19%) from the previous two

experiments is obtained. This experiment is by far better than those two experiments done without boosting.

The last experiment is again a similar boosting experiment done by using PART rule induction algorithm. But here, the changed parameter is the pruning state, which is made to be "TRUE", i.e. pruning do not happen during model development. This experiment has performed well in all parameters except with the number of rules generated.

Therefore, from the above four experiments we can understand that the two boosted experiments have good predictive accuracy than those done using the base classifier. The overall performance of each experiment is presented in the table below.

Experiment	Accuracy	No. of rules	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
PART pruned	58.41%	1,477	0.584	0.208	0.577	0.58	0.851
PART unprunned	61.56%	1,711	0.616	0.193	0.611	0.612	0.878
AdaBoostM1 + Pruned PART	81.19%	573	0.812	0.095	0.811	0.811	0.943
AdaBoostM1 + Unprunned PART	**82.01%**	843	**0.82**	**0.089**	**0.82**	**0.82**	**0.947**

Table 5.3: PART Experiments Performance Evaluation for the Sixth month CD4 count

In reference to the above summary, the unprunned boosted PART experiment has achieved a better performance in terms of accuracy, TP Rate (Sensitivity), FP Rate and the area under the ROC curve. The slight difference it has on the parameter number of rules in comparison with the boosted and pruned PART has no influence on the selection of the final better model. Thus, the unprunned boosted PART has a better performance in all the parameters except the number of rules generated. Consequently, the unprunned boosted PART experiment has been selected to be compared with the other classifier models developed in this experimental category. A table presenting improvements seen in the classifier's predictive performance over the minority classes by comparing the boosted unprunned PART and the unboosted one is appended in *Appendix-D*.

As indicated in the table under *Appendix- D*, all the evaluation parameter values have increased both in majority and minority classes for the boosted model. But the intention of boosting is improving the classifier's ability of predicting instances under the minority classes better than the model made only by the base classifier. Accordingly, in the two minority classes ("**350 - 499**" and **">=500**") the improvement is clearly indicated.

5.2.3. Sequential Minimal Optimization (SMO) Experiments

Two experiments are conducted using sequential minimal optimization algorithm by taking the default parameter values and the boosting algorithm. Here boosting is also used as a separate experimental scenario taking SMO with its default values. Unlike experiments done above, the boosting experiment here is failed to boost the TPR of minority classes. The detailed summary of the findings of each experiment is presented in the table below.

Experiment	Accuracy	CCI	ICI	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
SMO with default parameters	53.78%	3,900	3,352	0.538	0.315	0.409	0.462	**0.668**
AdaBoostM1 + SMO with default parameters	53.78%	3,900	3,352	0.538	0.315	0.409	0.462	**0.610**

Table 5.4: SMO Experiments Performance Evaluation for the Sixth Month CD4 Count

The summary depicts that sequential minimal optimization techniques are not good at developing classifiers for multiple class outcome variables. The result obtained from this experimentation is not used to be compared with other classifiers developed in this category of experimentation due to its poor performance.

5.2.4. Multilayer Perceptron (MLP) Experiments

Multilayer Perceptron is a neural network based classification algorithm which uses back propagation to classify instances in to known classes. Like the SMO experiment done above, here also two experiments are done by taking the default parameters values internal to the algorithm and using the boosting algorithm while the settings for MLP remains the same. As it can be seen in the table below, the first experiment has recorded an accuracy level of 59.97% and

ROC area of 0.765. In the second boosted experiment, the accuracy has increased to 64.96% with a ROC area of 0.846. Therefore, the boosting algorithm has contributed some to adjust the classifier's predictive accuracy. But in contrast to the previously done J48 and PART experiments this model is not good enough to be taken as a better classifier. Therefore, this model is not selected for comparison with J48 and PART classifier models. The following table presents the overall summary of the parameters used in comparing the performance of the above two experiments.

Experiment	Accuracy	CCI	ICI	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
MLP with default parameters	59.97%	4,349	2,903	0.6	0.225	0.581	0.583	0.765
AdaBoostM1 with MLP	**64.96%**	**4,711**	**2,541**	**0.65**	**0.203**	**0.639**	**0.633**	**0.846**

Table 5.5: MLP Experiments Performance Evaluation for the sixth Month CD4 Count

5.2.5. Model Evaluation

The model selection is done based on the statistical summary obtained from the WEKA machine learning environment. The following parameters are selected to compare classifiers done using the four mining algorithms; mean absolute error, accuracy of the model, sensitivity (TPR), False Positive Rate, F-measure and area under the ROC curve.

The summarized view of the comparison is presented in the following table by addressing all the necessary parameters.

No.	Model	Accuracy	Precision	F-Measure	Mean Absolute Error	TPR	FPR	AUC
1	Boosted and unprunned J48	**82.10%**	0.821	0.821	0.0496	0.821	0.090	**0.945**
2	Boosted and unprunned PART	**82.01%**	0.820	0.820	**0.0482**	0.820	**0.089**	**0.947**

Table 5.6: Selected Models Comparison for the Sixth Month CD4 Count

Considering the above numerical values labeled under each of the evaluation parameters, the boosted and unprunned J48 model has revealed a better performance. Hence, accuracy only can't be a valid qualification criteria for imbalanced datasets, the mean absolute error which is the average error calculated on those tests during ten iterations is a very good criteria both from data mining and statistical perspective [25]. In addition to mean absolute error, AUC is also a recommended parameter to evaluate model performance in the case of such imbalanced dataset. The ROC area is the area indicating the proportion of true positives versus false positives by putting the TPR on the Y-axis and FPR on the x-axis. The higher the TPR and the lower the FPR indicates maximum ROC area which is again an indication of good classifier. Therefore, the unprunned and boosted PART model has revealed better performance in the above two parameters (Mean Absolute Error and AUC) even though the J48 counterpart has a bit advance over the parameter accuracy.

Therefore, taking the mean absolute error, FPR and ROC area; the **boosted and unprunned PART** has shown the best performance and has been selected as a best classifier model for the sixth month CD4 count. The Weka output of the model is appended in *Appendix E*.

5.2.6. Rules Generated from the selected Model

The rules are extracted from the selected unprunned and boosted PART classifier based on the number of instances classified under each of the four classes and the baseline CD4 count status.

Initially, a total of fifty-eight rules were identified by taking 67% correct classification rate of each rules as a selection criterion. In consideration to the above established criteria, fifteen best rules those which are achieving a better accuracy were selected by taking the baseline CD4 count and functional status as a reference. In addition to the accuracy of the rules, the relevance of each rule to the public health is affirmed by domain experts. Therefore, the following rules are selected to represent the knowledgebase of sixth month CD4 count model of ART started patient.

Rule #1: If Baseline CD4 Count < 50 and Original Regimen = "1-a" and Age = 25 - 29 and Pregnancy Status = NO and Marital Status = Never Married AND Functional Status = A, then sixth month CD4 count is expected to be below 200 (11.19/4.68).

Rule #2: If Baseline CD4 Count < 50 and Age = 35 - 39 and Original Regimen = "1-a" and Educational Status = Secondary and Functional Status = W AND Baseline WHO Stage = Three AND Sex = Male, then sixth month CD4 count is expected to be below 200. (17.71/8.66)

Rule #3: If Baseline CD4 Count = 50 - 99 and Pregnancy Status = NA and Age = 30 - 34 and Baseline WHO Stage = Three, then sixth month CD4 count is expected to be below **200**. (42.25/23.86)

Rule #4: If Baseline CD4 Count >= 200 and Educational Status = Secondary and Functional Status = "A" and Sex = Female and Original Regimen = "1-a", then sixth month CD4 count is expected to be below **200** (40.26/23.86).

Rule #5: If Baseline CD4 Count >= 200 and Educational Status = Primary and Age = 25 - 29 and Original Regimen = "1-a" and Marital Status = Married and Functional Status = W and Family Planning = YES AND Baseline WHO Stage = Three, then sixth month CD4 count is expected to be between **200 and 349**. (19.78/9.69)

Rule #6: If Baseline CD4 Count = 100 - 199 and Sex = Male and Marital Status = Married and Original Regimen = 1-e and Age = 40 - 44 AND Baseline WHO Stage = Three, then sixth month CD4 count is expected to be between **200 and 349**. (20.24/10.0)

Rule #7: If Baseline CD4 Count >= 200 and Educational Status = Tertiary AND Marital Status = Married AND Functional Status = W AND Original Regimen = 1-e AND Age = 30 – 34, then sixth month CD4 count is expected to be between **350 and 499**. (10.94/5.37)

Rule #8: If Baseline CD4 Count < 50 and Original Regimen = 1-a and Functional Status = A and Pregnancy Status = NO and Marital Status = Married and Baseline WHO Stage = Three, then sixth month CD4 count is expected to be below **200**. (4.87/0.96)

Rule #9: If Baseline CD4 Count = 100 - 199 and Marital Status = Never Married and Age = 18 - 24 and Family Planning = YES and Functional Status = A and Original Regimen = 1-a and Baseline WHO Stage = Three, then sixth month CD4 count is expected to be **above 500**. (104.98/66.03)

Rule #10: If Baseline CD4 Count = 50 - 99 and Family Planning = NO and Age = 30 - 34 and Pregnancy Status = NA and Marital Status = Married, then sixth month CD4 count is expected to be between **200 and 349**. (14.64/7.31)

Rule #11: If Baseline CD4 Count < 50 and Original Regimen = "1-a" and Age = 25 - 29 and Pregnancy Status = NO and Marital Status = Married AND Baseline WHO Stage = Three AND Family Planning = NO, then sixth month CD4 count is expected to be **below 200**. (15.18/7.5)

Rule #12: If Baseline CD4 Count < 50 and Original Regimen = 1-b and Sex = Male and Marital Status = Never Married and Age = 35 – 39, then sixth month CD4 count is expected to be between **200 and 349**. (10.78/4.85)

Rule #13: If BaselineCD4Count = 100 - 199 and Family Planning = NO and Marital Status = Married and Original Regimen = "1-a" and Baseline WHO Stage = Three and Age = 30 - 34 and Functional Status = A and Sex = Female, then sixth month CD4 count is expected to be below 200. (11.41/4.93)

Rule #14: If Baseline CD4 Count = 100 - 199 and Functional Status = W and Sex = Female and Original Regimen = 1-e and Marital Status = Married and Educational Status = Primary and Baseline WHO Stage = Two, then sixth month CD4 count is expected to be between 200 and 349. (16.72/6.97)

Rule #15: If Baseline CD4 Count = 50 - 99 and Educational Status = Primary and Original Regimen = "1-a" and Functional Status = W and Marital Status = Married and Sex = Male and Age = 30 - 34, then sixth month CD4 count is expected to be below 200. (12.08/5.96)

5.2.7. Analysis of the Selected Rules

The above rules are not the one and only rules generated from the selected classifier model, rather they constitute those rules which are selected based on having a higher number of correctly classified instances and significance to the public health.

Generally, the rules are categorized in to four block groups based on the CD4 count value taken when patients begin the therapy. Those who initiated the therapy with CD4 count of below 50 are less likely to increase their CD4 count than those above 50. It has been observed in the rules that male's show a very good progress in their CD4 counts than females after six months of therapy. This model has also revealed that, patients' aged above 40 during initiation of therapy are unlikely to regain the average normal CD4 counts of an HIV patient. The other significant variable to determine improvement over CD4 count is the use of different family planning methods. Accordingly, those using family planning methods are less vulnerable to decrement in CD4 counts.

Overall, patients beginning the therapy with a baseline count of below 50 are less likely that their CD4 count advances to the next higher classes of CD4 count than those beginning with a count above this threshold.

5.3. Experimentations to Model Twelfth Month CD4 Count

In this section of experimentation, all algorithms used for building a classifier model for the sixth month CD4 count are used in the presence of the entire dataset to train, test and build classifiers. In addition to the ten predicting variables used in the previous experiments, the outcome variable "**CD4 count after sixth months**" is also used as a predicting variable in all experiments done here. Therefore, the next experiments are done using these eleven predicting variables.

5.3.1. J48 Experiments

Four experiments are done using J48 decision tree classifier. The confidence factor is also the same with the previous J48 experiments which is **0.5**. The experimental settings are:

Setting #1: J48 Experiment with pruning

Setting #2: J48 Experiment without pruning

Setting #3: AdaBoostM1 Experiment with its default parameters and taking J48 with pruning as a base classifier

Setting #4: AdaBoostM1 Experiment with its default parameters and taking J48 without pruning as a base classifier

In the first scenario, the experiment has generated a classifier having an accuracy of **66.85%** and a tree with a size of **4,605** and **3,516** leaves. This is relatively the least accuracy level scored in this category with a complex tree structure.

In the second experiment, a relatively better classifier with an accuracy of **72.48%** and a tree with a size of **4,844** and **3,696** leaves is generated. This experiment has produced relatively larger tree with increased number of leafs.

In the third experiment, a classifier with an accuracy of 89.81% which is better than the previous classifiers developed using the base classifier. Here the boosting algorithm has scaled up the accuracy from 68.39% (Pruned J48) to this higher value and also the tree size and number of leafs are lowered down to a less complex tree.

The fourth experiment is the boosted version of the second experiment (Unprunned J48 experiment), where a classifier with best performance relative to the previous three experiments is built. This classifier scored an accuracy of 89.96% with a tree of size 1,868 and 1,469 leaves. The detailed summary of each experiment is presented in the table below.

Experiment	Accuracy	Tree Size	Leaf Size	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
J48 pruned	68.39%	4,605	3,516	0.684	0.159	0.683	0.684	0.907
J48 unprunned	72.48%	4,844	3,696	0.725	0.134	0.725	0.725	0.928
AdaBoostM1 +Pruned J48	89.81%	1,304	1,010	0.898	0.051	0.898	0.898	0.971
AdaBoostM1 + Unprunned J48	**89.96%**	1,868	1,469	**0.9**	**0.05**	**0.9**	**0.9**	**0.974**

Table 5.7: J48 Experiments Performance Evaluation for the Twelfth Month CD4 Count

From the above data presented in the table, taking the accuracy level and less complexity of the decision tree, the boosted experiments are selected for further comparisons. Accordingly, the unprunned and boosted J48 classifier has higher classifier accuracy than that of the pruned one. In addition to this, the AUC is also higher for the unprunned and boosted J48 classifier than the pruned one. The only parameter where the pruned and boosted J48 exceeded the unprunned one is on the size of tree and number of leafs generated. Therefore, considering the weighted performance parameters and accuracy; the **boosted and unprunned J48** classifier has been selected to be compared with those classifiers generated using other classification algorithms.

5.3.2. PART Experiments

Like the J48 experiments done above in this category of experimentation, here also besides the ten predicting variables the outcome variable "**CD4 Count after Six Months**" is also used as a predicting variable to build a classifier for the twelfth month CD4 count. Moreover, the confidence factor used in the previous experiments is also maintained here. Accordingly, four experiments are conducted based on the following experimental scenarios.

Setting #1: PART Experiment with pruning.

Setting #2: PART Experiment without pruning.

Setting #3: AdaBoostM1 Experiment with its default parameters and taking PART with pruning as a base classifier.

Setting #4: AdaBoostM1 Experiment with its default parameters and taking PART without pruning as a base classifier.

In the first setting of the PART experiment, a classifier with an accuracy of **65.04%** and **1,453** rules is obtained.

In the second experimental setting with the only change of the parameter of pruning to not prune the tree, a model with accuracy of **71.77%** and **1,538** rules is constructed.

In the third experiment, a boosting algorithm (AdaBoostM1) is used together with the base classifier PART while the pruning parameter is turned on. The experiment has generated a classifier with an accuracy of **89.81%** with **523** decision rules.

In the fourth experiment, again boosting algorithm together with unprunned PART classifier has generated a model having an accuracy of **90.29%** and **753** rules. Therefore, these two boosted experiments have scored better classifier accuracy than the base classifiers.

The detail of the each of these four experimental results is presented in the following table.

Experiment	**Accuracy**	**Number of rules**	**WTP Rate**	**WFP Rate**	**WPrecision**	**WF-measure**	**WROC Area**
PART pruned	65.04%	1,453	0.65	0.172	0.65	0.65	0.886
PART unprunned	71.77%	1,538	0.718	0.132	0.72	0.719	0.922
AdaBoostM1 + Pruned PART	89.81%	523	0.898	0.051	0.898	0.898	0.970
AdaBoostM1 + Unprunned PART	**90.29%**	753	**0.903**	**0.05**	**0.903**	**0.903**	**0.973**

Table 5.8: PART Experiments Performance Evaluation for the Twelfth Month CD4 Count

Hence, the boosted experiments have shown a better performance over the base classifiers, the model selection comparison is totally focused on the two boosted experiments.

The unprunned and boosted PART experiment has attained an accuracy level higher than the boosted and pruned one. Besides to accuracy, this model also scored better results in the other weighted parameters; TPR, FPR, F-measure, Precision and AUC. Thus, taking the achieved accuracy level and the other weighed parameters, **boosted and unprunned PART** classifier model has been selected to be compared with other models generated in this experimental category.

5.3.3. Sequential Minimal Optimization (SMO) Experiments

As it has been done for the sixth month CD4 count classifier modeling using SMO, the same trend is also followed here to develop the classifiers. This indicates that two experiments of SMO with the default parameters and with the addition of boosting algorithm are conducted to select the better model.

Accordingly, in the first experiment a classifier with an accuracy of 56.73% and with a weighted ROC area of 0.729 is generated. In the second boosting intercepted experiment, a classifier with similar accuracy with the base classifier and with a weighted reduced ROC area of 0.689 is constructed. The detail of the classifiers is tabulated in the following way.

Experiment	Accuracy	CCI	ICI	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
SMO with default parameters	56.73%	4,114	3,138	0.567	0.278	0.560	0.544	0.729
AdaBoostM1 + SMO with default parameters	56.73%	4,114	3,138	0.567	0.278	0.560	0.544	0.689

Table 5.9: SMO Experiments Performance Evaluation for the Twelfth Month CD4 Count

From the above summary it is easy to understand that, the boosting algorithm did not brought any change on the outcome recorded due to the base classifier. Like similar experiments done for the six month CD4 count prediction by using this algorithm, here also this model is not selected to be compared with other models in this category of experimentation for similar reason.

5.3.4. Multilayer Perceptron (MLP) Experiments

Two experiments considering the default values of the parameters and another one by using a boosting algorithm is conducted to determine whether artificial neural network is good for multiclass outcome variable or not.

Accordingly, the first experiment revealed a classifier with an accuracy of 66.26%. The second experiment has generated a better classifier with relatively higher accuracy of 80.53% and a better coverage of area under the ROC curve. Therefore, from these two experiments the second one is taken for a comparison with other classifier models built in this category of experiments.

Therefore, the experiment supported with AdaBoostM1 is selected to be compared with previously generated classifiers. The following table indicates the overall performance of the two MLP experiments.

Experiment	Accuracy	CCI	ICI	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
MLP with default parameters	66.26%	4,805	2,447	0.663	0.191	0.662	0.659	0.811
AdaBoostM1 + MLP with default parameters	**80.53%**	**5,840**	**1,412**	**0.805**	**0.113**	**0.807**	**0.804**	**0.942**

Table 5.10: MLP Experiments Performance Evaluation for the Twelfth Month CD4 Count

5.3.5. Model Evaluation

Model selection for these set of experiments is done following similar selection parameters used for the previous outcome variable. Similarly, classifier's accuracy, weighted precision, weighted TPR, weighted FPR, weighted F-measure, mean absolute error and area under the ROC curve are used to select the best model among those selected in the previous steps.

The detailed summary on the performance of the three selected classifiers is presented in the table below.

No.	Model	Accuracy	Precision	F-Measure	Mean Absolute Error	TPR	FPR	AUC
1	Boosted and unprunned J48 classifier	89.96%	0.900	0.900	0.0496	0.900	0.05	0.974
2	Boosted and unprunned PART classifier	**90.29%**	**0.903**	**0.903**	**0.0482**	**0.903**	0.05	**0.973**
3	Boosted MLP Classifier	80.53%	0.807	0.804	0.1434	0.805	0.113	0.942

Table 5.11: Selected Classifier Models Comparison for the Twelfth Month CD4 Count

As it can be seen in the above table, the boosted and unprunned PART classifier has a lead on both boosted and unprunned J48 and boosted MLP classifiers in almost all of the presented parameters. The only parameter where the J48 classifier has got an advance over PART is the AUC with a very small (**0.001**) addition. Therefore, considering the accuracy, mean absolute error, TPR, FPR, Precision and F-measure the PART classifier is selected as a model to predict twelfth month CD4 count and the associated rules are generated from this classifier. The sample Weka output view of the selected model is appended in *Appendix F*.

5.3.6. Rules Generated from the selected Model

In this category of experimentation the final selected model which is the Boosted and unprunned PART classifier is used to generate rules relevant to the domain. The rules are extracted following similar selection criteria used for the sixth month CD4 counts predictive model. Accordingly, four sets of rules were selected i.e. a category formed based on the sixth month CD4 count status less than 200, between 200 and 349, between 350 and 499 and greater than or equal to 500. Consequently, a total of sixty-seven rules were selected based on the number of instances classified by those rules. Fifteen best rules were selected from each of the four classes based on the accuracy of rules. The following are selected best rules generated from the identified model.

Rule #1: If CD4 Count After Six Months < 200 and Functional Status = A and Original Regimen = 1-a and Educational Status = No Education, then twelfth month CD4 count is expected to be below 200. (5.55/0.99)

Rule #2: If CD4 Count After Six Months < 200 and Functional Status = A and Original Regimen = 1-a and Baseline CD4 Count = 50 - 99 and Educational Status = Secondary and Marital Status = Never Married, then twelfth month CD4 count is expected to be between 200 and 349. (33.77/16.71)

Rule #3: If CD4 Count After Six Months < 200 and Marital Status = Never Married and Baseline CD4 Count = 50 - 99 and Functional Status = W and Educational Status = Secondary and Sex = Male and Age = 30 - 34, then twelfth month CD4 count is expected to be below 200. (44.85/20.01)

Rule #4: If CD4 Count After Six Months < 200 and Functional Status = A and Original Regimen = 1-a and Marital Status = Never Married and Educational Status = Secondary and Age = 18 - 24, then twelfth month CD4 count is expected to be below 200. (47.68/20.39)

Rule #5: If CD4 Count After Six Months = 200 - 349 and Educational Status = Primary and Functional Status = W and Marital Status = Married and Sex = Female and Baseline WHO Stage = Three and Age = 25 - 29 and BaselineCD4Count >= 200, then twelfth month CD4 count is expected to be below 200. (56.72/23.25)

Rule #6: If CD4 Count After Six Months = 200 - 349 and Baseline CD4 Count = 100 - 199 and Functional Status = W and Baseline WHO Stage = Three and Age = 40 - 44 and Marital Status = Married and Original Regimen = 1-e, then twelfth month CD4 count is expected to be between 200 and 349. (46.89/19.58)

Rule #7: If CD4 Count After Six Months = 350 - 499 AND Baseline CD4 Count >= 200 and Original Regimen = 1-e and Baseline WHO Stage = Two and Sex = Male and Age = 30 - 34, then twelfth month CD4 count is expected to be between 200 and 349. (31.44/15.7)

Rule #8: If CD4 Count After Six Months < 200 and Functional Status = A and Marital Status = Married and Original Regimen = 1-a and Baseline WHO Stage = Three and Baseline CD4 Count < 50 and Age = 35 - 39 AND Family Planning = NO, then twelfth month CD4 count is expected to be below 200. (46.49/20.01)

Rule #9: If CD4 Count After Six Months >=500 and Original Regimen = 1-a and Family Planning = YES and Educational Status = Secondary and Age = 18 - 24 and Baseline CD4 Count = 100 - 199, then twelfth month CD4 count is expected to be above 500. (31.44/15.7)

Rule #10: If CD4 Count After Six Months = 200 - 349 and Educational Status = Secondary and Functional Status = W and Baseline CD4 Count = 50 - 99 and Family Planning = YES and Baseline WHO Stage = Three and Original Regimen = 1-b, then twelfth month CD4 count is expected to be between 200 and 349. (33.96/16.94)

Rule #11: If CD4 Count After Six Months = 200 - 349 and Marital Status = Married and Original Regimen = 1-b and Family Planning = YES and Educational Status = Secondary and Age = 18 - 24 and Sex = Female, then twelfth month CD4 count is expected to be between 350 and 499. (31.48/15.0)

Rule #12: If CD4 Count After Six Months = 200 - 349 and Educational Status = Primary and Baseline WHO Stage = Two and Original Regimen = 1-a and Marital Status = Married and Baseline CD4 Count = 100 - 199 and Family Planning = YES, then twelfth month CD4 count is expected to be between 350 and 499. (33.86/16.5)

Rule #13: If CD4 Count After Six Months = 200 - 349 and Baseline CD4 Count = 100 - 199 and Functional Status = W and Baseline WHO Stage = Three and Age = 40 - 44 and Marital Status = Married and Educational Status = Secondary, then twelfth month CD4 count is expected to be between 350 and 499. (47.04/20.57)

Rule #14: If CD4 Count After Six Months < 200 and Baseline CD4 Count = 50 - 99 and Sex = Male and Baseline WHO Stage = Three and Original Regimen = 1-a and Age = 35 - 39, then twelfth month CD4 count is expected to be below 200. (47.54/19.59)

Rule #15: If CD4 Count After Six Months = 350 - 499 and Baseline CD4 Count >= 200 and Functional Status = W and Sex = Female and Original Regimen = 1-a and Educational Status = Secondary and Age = 18 - 24, then twelfth month CD4 count is expected to be between 350 and 499. (57.02/22.15)

5.3.7. Analysis of the Selected Rules

A total of fifteen rules and ten out of this have matching rules within the previous six month rules are selected to model the twelfth month CD4 counts of a patient. Here, the previous records were used to select these rules based on relevant and meaningful overlapping observed with previous rules. Accordingly, those rules which overlap with previous rules and those suggested by the domain experts are included. In addition to these ten rules, five rules which do not have a match with the sixth month rules are selected again based on the comment received from domain experts.

In the rules it is clearly visible that, those patients having a CD4 count above 200 during the previous sixth month count has an increasing or sustained improvement on their CD4 count. But those with baseline CD4 count below 50 and six month CD4 counts below 200 are expected to have a CD4 count below 200 again after twelve months of therapy with exceptions of married patients.

As it has been presented in **Section 4.3.7**, aging here is also associated with poor CD4 counts. Those with age range below 40 have shown improvements in their number of CD4 cells given that they have scored a better CD4 counts at the six month.

Marital status and Educational level has influence on the improvement of CD4 count. Those who are married and better educated are associated with positive increment than others. This coincides with studies done by the FHAPCO.

These findings are validated by expert from FHAPCO and with the findings available on a document prepared by this organization on ART Scale-up in Ethiopia [72].

5.4. Experimentations to Model Eighteenth Month CD4 Count

The experimental setup arranged to build up a model for the eighteenth month CD4 count is similar with the previous experimentations except adjustments made on the number of predicting variables used. Here, the sixth and twelfth month CD4 counts are included in to the category of predicting variables to forecast the eighteenth month CD4 count forming a total of twelve predicting variables. The detail of experiments conducted using each of the four classifier algorithms together with the boosting algorithm and the associated discussions over each experimental result is presented in the next sub sections.

5.4.1. J48 Experiments

Experimental settings followed in the above two sections are also applied here to conduct the J48 experiments. Accordingly the following four experimental settings were identified to go through the experimentation.

Setting #1: J48 Experiment with pruning

Setting #2: J48 Experiment without pruning

Setting #3: AdaBoostM1 Experiment with its default parameters and taking J48 with pruning as a base classifier

Setting #4: AdaBoostM1 Experiment with its default parameters and taking J48 without pruning as a base classifier

In the first experiment a classifier with an accuracy calculated to be **72.70%** and a tree with a size of **4,236** and **3,308** leaves is built taking the default parameters of the algorithm and confidence factor (CF = 0.5).

In the second experiment the accuracy has grown a little bit higher to **77.94%** which is the maximum accuracy achieved among the experiments done using J48 for the three outcome variables under normal conditions. The tree size has also grown to **4,668** with **3,650** leaves.

In the third experiment boosting algorithm is applied on the pruned J48 decision tree classifier. Consequently, a model with an accuracy of **91.95%** and a decision tree with a size of **1,226** and **960** leaves is generated.

In the last scenario of the experimentation similar boosting algorithm is applied to the unprunned J48 decision tree classifier. The resulted model has yielded a better classifier with accuracy of **91.70%.**

The overall summary indicating the performance of each classifier is presented in the next table.

Experiment	Accuracy	Tree Size	Leaf Size	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
J48 pruned	72.70%	4,236	3,308	0.727	0.127	0.727	0.727	0.925
J48 unprunned	77.94%	4,668	3,650	0.779	0.101	0.780	0.780	0.945
AdaBoostM1 with pruned J48	91.95%	1,226	960	0.919	0.039	0.920	0.919	0.977
AdaBoostM1 with unprunned J48	91.70%	1,655	1,301	0.917	0.040	0.917	0.917	0.979

Table 5.12: J48 Experiments Performance Evaluation for the Eighteenth Month CD4 Count

In these four experiments, it is very clear that each of the comparing parameters are represented with figures indicating the performance level of each classifier. Accordingly, the boosted classifier model has performed better than the base classifier. Therefore, the two boosted J48 classifier models are compared in order to select the final best model that can be compared with classifiers generated by other algorithms.

The pruned and boosted J48 algorithm has generated relatively less complex decision tree in comparison with the unprunned and boosted one. And also, this classifier has better accuracy, TPR, FPR, Precision and F-measure values than the boosted and unprunned J48 model. Even though the unprunned and boosted J48 has relatively lower accuracy and those mentioned parameter values, it has a relatively higher AUC value than its counterpart. Therefore, principally taking the parameters "Accuracy" and "AUC", the **boosted and unprunned J48** classifier model has been selected to be compared with other classifier models built in this category.

5.4.2. PART Experiments

Similar to the above J48 experiment, all experiments are done by using twelve predicting attributes including previously used outcome variables (six month CD4 count and twelfth month

CD4 count).The experiments are conducted according to the following four experimental scenarios.

Setting #1: PART Experiment with pruning.

Setting #2: PART Experiment without pruning.

Setting #3: AdaBoostM1 Experiment with its default parameters and taking PART with pruning as a base classifier.

Setting #4: AdaBoostM1 Experiment with its default parameters and taking PART without pruning as a base classifier.

In the first experiment a decision list containing **1,269** rules is generated with classifier's accuracy of **70.12%**. This indicates only **29.88%** instances were wrongly classified among the four classes.

In the second experiment **1,379** rules are generated. In contrast with the first experiment, this experiment has developed a classifier with an accuracy of **77.83%** which means only **25.76%** of the instances were wrongly classified.

In the third experiment the classifier is boosted using the boosting algorithm AdaBoostM1 and resulted in building better classifier. Here, the accuracy is reached to **91.85%** which is a maximum of the experiments done in this category. The number of rules also minimum when compared with the three classifiers generated here.

In the last experiment a relatively better accuracy level is reached which is better than the base classifiers even though it is a bit less than the third experiment.

The detail of the comparison made on these four experiments is presented in the table below.

Experiment	Accuracy	Number of Rules	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
PART pruned	70.12%	1,269	0.701	0.138	0.701	0.701	0.910
PART unprunned	77.83%	1,379	0.778	0.1	0.78	0.779	0.939
AdaBoostM1 with pruned PART	**91.85%**	**456**	**0.919**	**0.039**	**0.919**	**0.919**	**0.979**
AdaBoostM1 with unprunned PART	91.45%	636	0.915	0.04	0.915	0.915	0.979

Table 5.13: PART Experiments Performance Evaluation for the Eighteenth month CD4 Count

In the above four experiments the accuracy has increased in a very large gap from the base classifiers to the boosted ones. Therefore, the evaluation is totally targeted on the two boosted classifier models. The boosted and pruned PART classifier has better values in all parameters except the AUC value which is similar with that of the unprunned one. This indicates the **pruned and boosted PART** classifier is better than the unprunned one in this experimental scenario. Thus, the pruned and boosted PART classifier has been selected to be compared with other classifiers developed by other algorithms under this section.

5.4.3. Sequential Minimal Optimization (SMO) Experiments

Here also two experiments are undertaken by using the default parameters of SMO algorithm together with the boosting algorithm. The detailed summary of the experiments is presented in the following table.

Experiment	Accuracy	CCI	ICI	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
SMO with default parameters	60.73%	4,404	2,848	0.607	0.201	0.617	0.601	**0.771**
AdaBoostM1 + SMO with default parameters	60.73%	4,404	2,848	0.607	0.201	0.617	0.601	**0.762**

Table 5.14: SMO Experiments Performance Evaluation for the Eighteenth Month CD4 Count

Like similar SMO experiments done in the above sections, here also the boosting algorithm did not worked out to boost the performance of the base classifier. Therefore, the models found from these two experiments did not get compared with others in this group of experiments due to their poor performance. Generally, from what has been seen in these three boosting experiments, boosting using SMO as a base classifier is not good enough for classification tasks involving multiple classes and imbalanced dataset.

5.4.4. Multilayer Perceptron (MLP) Experiments

This is the last experiment conducted both under this category and from the entire set of experimentation. Similarly two MLP experiments, one with the default parameters and other with the integration of boosting algorithm on the previous setup is conducted to come up with a

better classifier model. As a result, the first experiment has generated a model with an accuracy of 72.49%, whereas the boosted experiment has generated a better classifier with an accuracy of 85.05%. The overall summary of the parameters used to rate the two experiments are presented in the following table.

Experiment	Accuracy	CCI	ICI	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
MLP with default parameters	72.49%	5,257	1,995	0.725	0.137	0.727	0.724	0.862
AdaBoostM1 + MLP with default parameters	**85.05%**	**6,168**	**1,084**	**0.851**	**0.078**	**0.853**	**0.851**	**0.960**

Table 5.15: MLP Experiments Performance Evaluation for the Eighteenth Month CD4 Count

Therefore, the boosted multilayer Perceptron is selected to be compared with other classifier model generated in this category.

5.4.5. Model Evaluation

Following similar evaluation techniques followed in the above sections, the best classifier model for the eighteenth month CD4 count is selected. Three selected models are compared to select the final best model of this scenario. Accordingly, the overall performance of each of the three classifier models is presented in the following table.

No.	Model	Accuracy	Precision	F-Measure	Mean Absolute Error	TPR	FPR	AUC
1	Boosted and unprunned J48	91.70%	0.917	0.917	0.0387	0.917	0.040	0.979
2	Boosted and pruned PART	**91.85%**	**0.919**	**0.919**	**0.0384**	**0.919**	**0.039**	**0.979**
3	Boosted MLP	85.05%	0.853	0.851	0.1027	0.851	0.078	0.960

Table 5.16: Selected Models Comparison for the Eighteenth Month CD4 Count

From the above table it is straightforward to pinpoint a better performing classifier model taking the presented parameter values. As a result, the boosted and pruned PART classifier model is selected as a best model of eighteenth month CD4 count by looking the parameters accuracy,

precision, F-measure, mean absolute error and TPR, even though similar AUC values are scored with the J48 classifier. The sample Weka output of the selected model is appended in *Appendix-G*.

5.4.6. Rules Generated from the selected Model

As it has been clearly presented above, the boosted and pruned PART classifier model is selected to extract the necessary knowledge (if-then rules) to apply for future new instances. Following this, interesting rules were extracted by setting a minimum number of correctly classified instances for each of the rules identified. Initially, thirty instances were set as a bench mark to select rules from the selected decision list so that seventy-seven interesting rules were selected under each of the four classes for further screening. The final fifteen rules were selected by taking those rules having a maximum number of correctly classified instances and got approval from domain experts and documents cited.

Rule #1: If CD4 Count After Twelve Months < 200 and Baseline WHO Stage = Two and Marital Status = Never Married, then eighteenth month CD4 count is expected to be below 200. (9.16/1.87)

Rule #2: If CD4 Count After Twelve Months < 200 and Family Planning = NO and Baseline WHO Stage = Three and CD4 Count After Six Months < 200 and Educational Status = Secondary and Original Regimen = 1-a and Marital Status = Married and Sex = Male, then eighteenth month CD4 count is between 200 and 349. (57.08/25.12)

Rule #3: If CD4Count After Twelve Months < 200 and Baseline WHO Stage = Two and Marital Status = Never Married, then eighteenth month CD4 count is expected to be below 200. (9.16/1.87)

Rule #4: CD4 Count After Twelve Months = 200 - 349 and CD4 Count After Six Months = 200 - 349 and Baseline WHO Stage = Three and Baseline CD4 Count = 100 - 199 and Family Planning = YES and Marital Status = Married and Original Regimen = 1-a, then eighteenth month CD4 count is between 200 and 349. (428.43/194.42)

Rule #5: If CD4 Count After Twelve Months = 200 - 349 and Baseline WHO Stage = Three and CD4 Count After Six Months = 200 - 349 and Pregnancy Status = NA and Baseline CD4 Count = 100 - 199 and Family Planning = YES, then eighteenth month CD4 count is between 350 and 499. (10.16/2.11)

Rule #6: If CD4 Count After Twelve Months = 200 - 349 and Sex = Female and CD4 Count After Six Months = 200 - 349 and Baseline WHO Stage = Three and Marital Status = Married and Age = 30 - 34 and Original Regimen = 1-e, then eighteenth month CD4 count is between 350 and 499. (55.53/25.12)

Rule #7: If CD4 Count After Twelve Months < 200 and CD4CountAfterSixMonths < 200 and Baseline WHO Stage = Three and BaselineCD4Count < 50 and Marital Status = Married and Educational Status = Primary and Functional Status = A, then eighteenth month CD4 count is below 200. (6.86/1.38)

Rule #8: If CD4 Count After Twelve Months >=500 and Age = 18 - 24 and Family Planning = YES and Marital Status = Married and Functional Status = W, then eighteenth month CD4 count is above 500. (90.98/30.92)

Rule #9: If CD4 Count After Twelve Months = 200 - 349 and Educational Status = Secondary and Marital Status = Married and CD4 Count After Six Months = 200 - 349 and Age = 30 - 34 and Baseline WHO Stage = Three and Family Planning = NO, then eighteenth month CD4 count is between 200 and 349. (55.7/25.29)

Rule #10: If CD4 Count After Twelve Months = 200 - 349 and CD4 Count After Six Months = 200 - 349 and Pregnancy Status = NO and Baseline WHO Stage = Three and Marital Status = Married and Educational Status = Secondary and Functional Status = W and Baseline CD4 Count >= 200 and Original Regimen = 1-a, then eighteenth month CD4 count is expected to be between 350 and 499. (138.88/56.04)

Rule #11: If CD4 Count After Twelve Months = 350 - 499 and Marital Status = Married and Family Planning = YES and Baseline CD4 Count >= 200 and Functional Status = W and Age = 30 - 34 and Educational Status = Secondary and Baseline WHO Stage = Three and Original Regimen = 1-b, then eighteenth month CD4 count is expected to be between 350 and 499. (82.15/30.92)

Rule #12: If CD4 Count After Twelve Months = 350 - 499 and Family Planning = YES and Marital Status = Married and CD4 Count After Six Months = 350 - 499 and Age = 25 - 29 and Educational Status = Secondary and Original Regimen = 1-a and Baseline CD4 Count = 100 - 199, then eighteenth month CD4 count is expected to be above 500. (55.53/25.12)

Rule #13: If CD4 Count After Twelve Months = 350 - 499 and Marital Status = Married and Family Planning = YES and CD4 Count After Six Months = 350 - 499 and Functional Status = B AND Educational Status = Tertiary, then eighteenth month CD4 count is expected to be above 500. (55.53/25.12)

Rule #14: If CD4 Count After Twelve Months = 350 - 499 and Family Planning = YES and Age = 25 - 29 and Educational Status = Secondary and CD4 Count After Six Months = 200 - 349 and Original Regimen = 1-a, then eighteenth month CD4 count is between 350 and 499. (15.32/6.2)

Rule #15: If CD4 Count After Twelve Months = 200 - 349 and Age = 18 - 24 and Original Regimen = 1-a and CD4 Count After Six Months = 200 - 349 and Sex = Female and Baseline WHO Stage = Two, then eighteenth month CD4 count is between 350 and 499. (13.09/3.17)

5.4.7. Analysis of the Selected Rules

The same story is also exhibited here like the previous rules seen in the sixth and twelfth month counts. Patients having a better CD4 count at sixth and twelfth month, most probably develop a very good immunity level as compared with others having poor count in one of those two previous counts. On the other hand if two of the previous counts are below 200, definitely the eighteenth month count is also expected to be below 200.

5.5. Discussion on Major Findings

5.5.1. Concatenated Rules of the three Models

The final knowledge or set of "if-then" rules are obtained by combining best performing rules from each of the three selected best models taking the overlapping class values. This is done so due to the reason that the output of the first outcome variable is feedback for the second and that of the second is feedback for the third. Hence, whenever a patient receives CD4 count during the sixth month visit, based on the result obtained, the physician provides counseling services, vitamin supplements, iron supplements and others to make the worst to be better. Therefore, to make the knowledge more realistic the rules are joined based on similar features contained in common in the three models and their predictive accuracy.

The process is started by first selecting best rules from the last outcome variable (CD4 count after eighteen months) and then looking back on those rules which coincides with the rules under the twelfth month CD4 count. Accordingly, those rules which are found in common are taken for the next comparison to be made with the six month CD4 count rules. Following the same procedure, the final best rules are selected in this final comparison made between the common rules extracted from twelfth and eighteenth month rules and the sixth month rules. The final integrated rules are constructed favoring more on those which seeks more attention during

therapy (i.e. those starting treatment at a baseline CD4 count below 50 and functional status ambulatory and bedridden) and those which have shown a dramatic improvement in the course of treatment. Therefore, these final rules contain predicting variables which are common to the three outcome variables. These unified rules are again used in prototyping the model to show the applicability of data mining for the identified domain.

The following eight rules are selected by joining selected rules from each of the three outcome variables. The detail of expert based judgment of the rules is given in the next sub section.

Rule #1: If Baseline CD4 Count < 50 and Original Regimen = 1-a and Age = 25 - 29 and Pregnancy Status = NO and Marital Status = Never Married and Functional Status = A and Educational Status = No Education and Baseline WHO Stage = Two and Sex = Female and Family Planning = No, then six month CD4 count is expected to be below 200, twelfth month CD4 count is expected to be below 200, and also eighteenth month CD4 count is expected to be below 200.

Rule #2: If Baseline CD4 Count = 50 - 99 and Pregnancy Status = NA and Age = 30 - 34 and Baseline WHO Stage = Three and Marital Status = Never Married and Functional Status = W and Educational Status = Secondary and Sex = Male and Family Planning = NO and Original Regimen = 1-a, then sixth month CD4 count is expected to be below 200, twelfth month CD4 count is expected to be below 200, and eighteenth month CD4 count is between 200 and 349.

Rule #3: If BaselineCD4Count >= 200 and Age = 18 – 24 and Educational Status = Secondary and Functional Status = A and Sex = Female and Pregnancy Status = No and Original Regimen = 1-a and Marital Status = Never Married AND Baseline WHO Stage = Two, then sixth month CD4 count is expected to be below 200; twelfth month CD4 count is expected to be below 200 and eighteenth month CD4 count is expected to be below 200.

Rule #4: If BaselineCD4Count >= 200 AND Age = 25 - 29 AND Original Regimen = 1-a AND Marital Status = Married AND Functional Status = W AND Family Planning = YES AND Baseline WHO Stage = Three AND Educational Status = Primary AND Sex = Female AND Pregnancy Status = No, then sixth month CD4 count is expected to be between 200 and 349, twelfth month CD4 count is expected to be below 200 and eighteenth month CD4 count is between 200 and 349.

Rule #5: If Baseline CD4 Count = 100 - 199 and Sex = Male and Marital Status = Married and Original Regimen = 1-e and Age = 40 - 44 and Baseline WHO Stage = Three and Functional Status = W and Pregnancy Status = NA and Family Planning = YES and Educational Status = Secondary, then sixth month CD4 count is expected to be between 200 and 349, twelfth month CD4 count is expected to be between 200 and 349, and eighteenth month CD4 count is also expected to be between 350 and 499.

Rule #6: If Baseline CD4 Count >= 200 and Educational Status = Tertiary and Marital Status = Married and Functional Status = W and Original Regimen = 1-e and Age = 30 – 34 and Baseline WHO Stage = Two and Sex = Male or Female, then sixth month CD4 count is expected to be between 350 and 499, twelfth month CD4 count is expected to be between 200 and 349 and eighteenth month CD4 count is expected to be between 350 and 499.

Rule #7: If Baseline CD4 Count < 50 and Original Regimen = 1-a and Functional Status = A and Pregnancy Status = NO and Marital Status = Married and Baseline WHO Stage = Three and Age = 35 - 39 and Sex = Female and Family Planning = NO and Educational Status = Primary, then sixth month CD4 count is expected to be below 200, twelfth month CD4 count is expected to be below 200, and eighteenth month CD4 count is also expected to be below 200.

Rule #8: If Baseline CD4 Count = 50 - 99 and Family Planning = NO and Age = 30 - 34 and Pregnancy Status = NA and Marital Status = Married and Educational Status = Secondary and Functional Status = W and Baseline WHO Stage = Three and Original Regimen = 1-b and Sex = Male, then sixth month CD4 count is expected to be between 200 and 349, twelfth month CD4 count is expected to be between 200 and 349, and also eighteenth month CD4 count is expected to be between 200 and 349.

5.5.2. Analysis of the Integrated Rules

The final rules are selected by taking the eighteenth month rules and finding their sole matches in rules under the twelfth month count. This process continues by taking those having an intersection from previous sets to find the final best rules by matching with the rules under the sixth month rules. Thus, a total of eight rules were identified to be important taking age, baseline CD4 count, functional status and previous CD4 count results as a criterion to select from the identified rules.

In these rules, we can see that those beginning ART with a CD4 count below 50 are less likely to develop a better immunity as compared to those beginning above 50 at any given age level. Those patients beginning the therapy at a baseline CD4 count of between hundred and two hundred, their CD4 count increases from time to time given that their marital status is married and they are using family planning methods. Whereas those beginning therapy at a baseline count of between 50 and 99, progressively their CD4 count advances to the levels where no opportunistic infections can occur. But those starting the therapy at a baseline count above two hundred, their immunity develop to a better level given that they are using different family planning techniques.

In general, a promising result is found in this research whereby facilities with no CD4 counting machine can make use of this knowledgebase to predict their patients immunological status based on the baseline clinical assessments and socio-demographic attributes. Also physicians can take a decision ahead before beginning a certain drug regimen or giving additional supplements or providing training for those who have shown no progress according to the model output.

5.5.3. Evaluation of the Discovered Knowledge

The discovered knowledge is presented in the form of if-then statements where experts in the area can make use of it taking the values of the ten predicting variables. In this study, two data mining goals were set to guide the overall flow of the study and reach on the final target. Accordingly, the first mining goal to be attained was: ***Given the socio-demographic data, baseline WHO clinical stage and baseline CD4 cells count, predict the CD4 cells count of a patient at six, twelve and eighteen months of therapy***. Let's take a patient record with the following details; ***Age = 26, Sex = Female, Marital Status = Never Married, Educational Status = No Education, Functional Status = Ambulatory, Drug Regimen = "1-a", Baseline WHO Stage = Two, Family Planning = No, Pregnancy Status = No and Baseline CD4 Count = 15***. Taking the first rule obtained after integration, the patient CD4 count after the sixth month is expected to be below 200; similarly the twelfth month also expected to be below 200; and finally at the eighteenth month its count also gets below 200. This indicates that a patient beginning with such socio-demographic, clinical and biological parameters needs a due attention to reverse the poor disease prognosis.

The other data mining goal was: ***From the identified predicting variables, determine those having a better prediction performance***. In this study, all the variables are selected based on the comments collected from the domain area experts and a review made on related literatures done in the area. Accordingly, all of the ten predicting variables are selected to enroll in the study. But during knowledge extraction some of the variables predominantly appeared in each of the rules while others occurred less frequently. Therefore, those which occurred frequently are taken as the most predicting variables than the others. Six variables are observed to be much more important in the identified if-then rules. These are; *Baseline CD4 counts, Original Regimen, Age, Family Planning usage Status, Functional Status*, and *Baseline WHO Stage*. But the rest,

Pregnancy Status, Marital Status, Sex, and Educational Status are used very less number of times, so that these variables have less predicting capability than the previous six ones.

5.6. Prototype Development

The last objective of this study was developing a prototype interface (system) that provides easy access to the identified knowledgebase. The final selected if-then rules are used to implement the selected best models from each of the three experimentations.

The programming tool used to host the identified rules is Microsoft visual basic 2008. Therefore, only those rules which are suggested to be important by domain experts are placed in to this prototype which means all the rules for predicting CD4 counts of a patient can't be answered by this prototype. The following picture is the main form used to run the commands to predict a patient's CD4 counts at sixth, twelfth and eighteenth month of therapy.

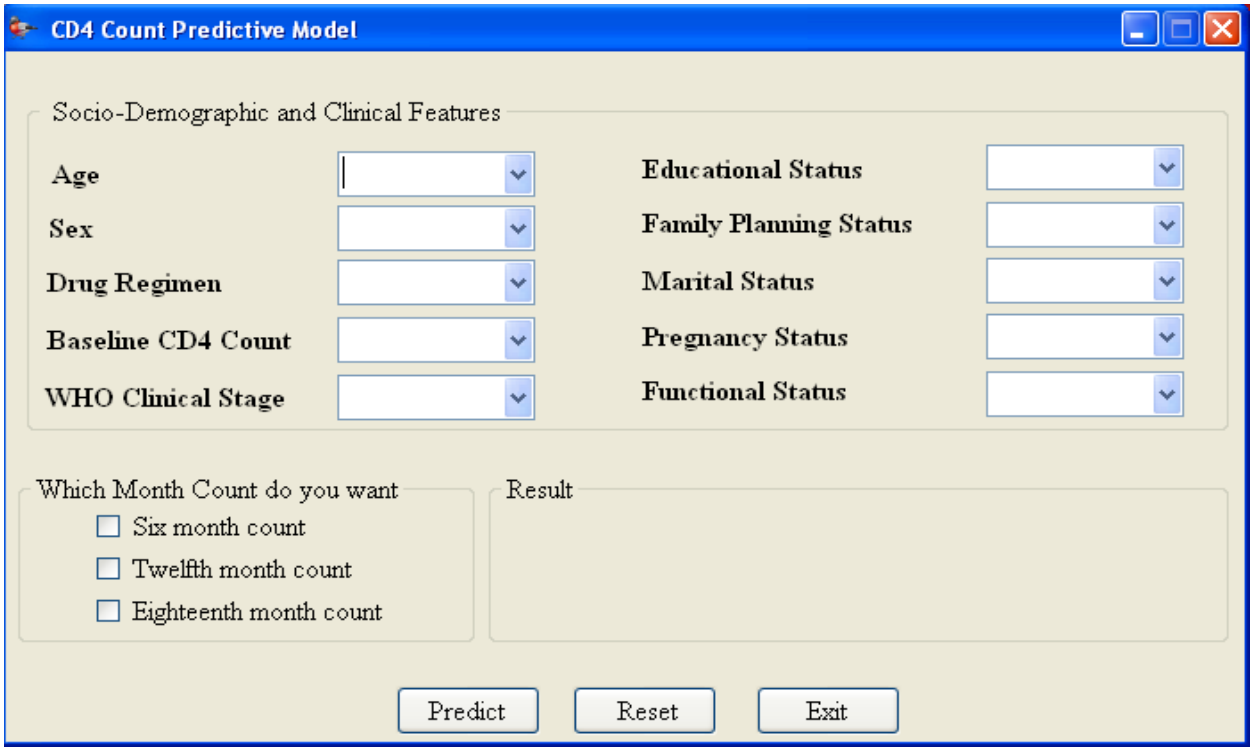

Figure 5.2: Main User Interface of the Prototype

CHAPTER SIX

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

6.1. Summary

Data mining is an intersection of artificial intelligence, machine learning, statistics, and database systems which targets to extract useful information from a data set and transform it into an understandable structure for further use [24]. Availability of large volume of data coupled with dramatic increase in the processing capability of computers created the best of environments for data mining applications [24].

The health sector is among those which are currently benefiting a lot from data mining by exploiting previously unknown or hidden knowledge from stored health records. Nowadays, ART datasets are growing at an alarming rate due to the availability of the service at no cost and in nearby health facilities [72]. The ART records contain patients demographic, clinical and biological features in addition to ART follow up information recorded to monitor the patient's progress over time.

This study was designed to predict the CD4 cells count of a patient after six, twelve and eighteenth months of therapy based on socio-demographic, biological and clinical features by using data mining techniques. To achieve this objective, related works in the area are thoroughly reviewed and domain experts are consulted in selecting the attributes vital for the study.

The data used for this study was collected from three government owned hospitals; namely Jimma, Bonga and Aman which are found in the south-western part of Ethiopia. Initially, 10,220 records were collected from the three hospitals. After making productive discussion with domain experts from those hospitals and FMOH, the final 7,252 records containing all the necessary instances were selected. Moreover, critical investigation of the business is done to understand the business and the data items contained in the dataset.

In order to get valid result or model of interest, the data had to undergo data preprocessing operation. Accordingly, data cleansing activities were performed i.e. handling outliers, inconsistent data, and missing values on the selected dataset. To support these tasks exploratory data analysis was applied both on the nominal and numeric attributes. Data reduction and

transformation strategies were also employed to make the data suitable for the selected data mining task. Overall, well prepared dataset was used to run the mining algorithms.

Totally, five mining algorithms were used in developing the classifier models. The four algorithms i.e. J48, PART, SMO and MLP are used as a base classifier and the other one i.e. AdaBoostM1 is used to boost the classifier's predictive accuracy. Better classifier accuracy is observed by integrating this ensemble learning technique in this study.

All predicting variables were included in each experimental scenarios conducted for the sixth month CD4 count. In the case of twelfth and eighteenth month CD4 count, the outcome variable used in the previous model was used as a predicting variable for the next consecutive model. Therefore, sixth month CD4 count was used as a predicting variable for the twelfth month CD4 count model and also sixth and twelfth month CD4 counts were used as a predicting variables for the eighteenth month CD4 count model.

Three major experimentations were conducted to build a model for the sixth, twelfth and eighteenth month CD4 counts. To come up with the final best model, the four classification mining algorithms together with the boosting algorithm were used in each experimental scenario. As a result, the boosted and unprunned PART classifier was selected for the sixth and twelfth month CD4 counts; whereas the boosted and pruned PART performed better from those experiments conducted for the eighteenth month CD4 counts.

Model comparison was made based on the performance evaluation parameters suggested for multi class problems with imbalanced dataset. Beside the usual accuracy, TP Rate and FP Rate; F-Measure and more importantly area under the ROC curve are used in comparing models for selection.

Finally, a prototype system (interface) of the identified knowledge was designed by picking overlapping rules from the three outcome variables to ease access by domain users.

Thus, findings and conclusions of this study are hoped to help physicians, patients, policy makers, hospitals, donors, government, and researchers to put the best direction for the successful implementation of ART programs in Ethiopia.

6.2. Conclusions

Based on a series of experiments conducted in the study, the following conclusions are made.

From classifier models generated for the sixth month CD4 count, the unprunned and boosted J48 classifier has attained a better accuracy but with lesser area under the ROC curve, and a bit higher mean absolute error and FPR than the boosted and unprunned PART model. This model has attained an accuracy level of **82.1%** with the area under ROC curve of **0.945** but the PART model has scored an accuracy of **81.01%** with AUC of **0.947**. Therefore, due to the better AUC and less mean absolute error the **unprunned and boosted PART** classifier model is selected as a best model for the twelfth month CD4 count.

From classifier models generated for the twelfth month CD4 count, again the **unprunned and boosted PART** classifier model has exceeded both the **J48** and **MLP** models selected for comparison due their better performance. From these three models PART classifier has achieved better performance in all the parameters used to rate the classifiers. Therefore, the boosted and unprunned PART classifier is selected to model the twelfth month CD4 count.

In the models generated for eighteenth month CD4 count, all of the three algorithms with the interference of boosting have exhibited a better performance than previously done experiments. Accordingly the **pruned PART** after boosting has yielded a better classifier with an accuracy of **91.85%** and a bit higher parameter values in all the rest except AUC than the J48 counterpart. The area under the ROC curve of both the PART and J48 experiments is equal (**0.975**). But PART has better performance indicators in all other parameters and thus it has been selected as a better classifier model for the eighteenth month CD4 count. The artificial neural network based classifier has scored a better performance at this stage but still it is lower than J48 and PART classifiers. Generally, ensemble learning techniques are very good approaches for bio-medical data sets with imbalanced multiple classes [70]. On the other hand support vector machine did not bring a better accuracy even after the induction of boosting algorithm in all of the three experimentations.

The finding of this study has indicated that *baseline CD4 count*, *original regimen*, *age*, *family planning usage status*, *functional status* and *baseline WHO clinical stage* are the most determining attributes to predict CD4 count of ART following patients. This findings are

approved by domain experts and literature cited on the success and challenges of ART programs in Ethiopia [72].

The findings indicated above entails that data mining techniques are suitable to develop predictive models using ART datasets.

6.3. Recommendations

The investigator suggests the following major recommendations for all concerned based on the observed findings.

Practitioners are the firsthand users of the identified knowledgebase. Therefore, they should take the initiative to experiment the model by predicting longitudinal CD4 counts of patients during initiation of therapy. This will help them to know the real application of the model to achieve the business objectives. This has a significant benefit to improve the patient's quality of life in general and; avoid unnecessary waiting time of patients to get their CD4 test result and wastage of resources while investigating CD4 counts.

The data managers or clerks are core to the quality of data. Therefore, they have to give due attention during data transcription in order to make it usable for the intended objective as well as to be used for researches claiming the dataset.

The hospital management should be positive in providing access to the medical data like the one used in this study without violating ethical issues to help researchers in finding valuable patterns residing in it. In addition to this, this management should arrange the necessary logistics essential for data transcription and keeping it safe from data errors. The findings in this study indicated the focus areas to work on so that the hospital management must plan activities accordingly.

Governmental and Non-governmental organization working on the provision of ART programs should use the findings of the study to identify the most disadvantaged group of the society who needs serious attention during the follow-up times. This further can help these organizations to identify the support areas for those identified focus groups and thereby address the goal for successful implementation of ART programs.

In addition to the above support, these bodies should also give attention to arranging trainings on data quality, privacy, and security management. Moreover, the government and those aid organizations should provide frequent trainings to data managers and clerks on data encoding

and on the essence of keeping quality data which can further be used for research purposes besides its immediate objective.

Researchers can make use of this thesis as a reference to conduct similar study by changing the study area, or by adding more instances to the existing dataset to validate the findings. The investigator also recommends researchers to try additional data mining tools like Rapid Miner and R besides Weka, due to recent advancements in machine learning tools especially those recommended here.

The investigator also recommends future works to use ensemble learning techniques in addition to the well known machine learning algorithms. Since biomedical data in general and ART datasets specifically are imbalanced and contain multiple classes [70]; this approach is the most recommended due to its capability to deal with such kind of datasets. Equally importantly attempting similar work by increasing the dataset could possibly yield a better classifier model.

Researches in the public health side can conduct clinical trials to compare the result obtained with the one done using data mining and recommend future directions over this scenario. In addition to this, a national based study needs to be conducted to make the model representativeness national and again its associated knowledge based system should be developed to use it at its full scale.

Finally but yet importantly, this thesis can be used for academic purposes being a reference for those looking to work in the identified problem domain or generally on data mining.

References

1. Steven JR, Thomas CQ, Chris B, Robert CB. Antiretroviral therapy where resources are limited. N Engl J Med; 2003; 348:1806-09.
2. Department of Health and Human Services. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Accessed on Dec 29, 2012. Available from: http://www.aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf.
3. WHO. Antiretroviral Therapy for HIV infection in adult's adolescents. 2nd Edition; 2006.
4. UNAIDS. World AIDS day report, 2011.
5. Federal HIV/AIDS Prevention and Control Office, Ministry of Health. Guidelines for implementation of the Antiretroviral Therapy in Ethiopia, July, 2007.
6. WHO. Antiretroviral Therapy [Internet]. Accessed on Dec 2, 2012. Available from: http://www.who.int/hiv/topics/treatment/art/en/index.html.
7. Gadelha A. Morbidity and survival in advanced AIDS in Rio Dejaneiro. Revised Ed. Trop Paulo; 2002.
8. Mellors. J, Munoz. A. Plasma viral load and CD4+lymphocytes as prognostic markers of HIV-1 infection. Ann Intern Med; 1997.
9. Behailu G. Constructing a Predictive Model for Determining CD4 Status of Patients Following Art: The Case of Jimma and Bonga Hospitals [Unpublished MSc Thesis]. Addis Ababa University: School of Information Science and School of Public Health; 2012.
10. Yashik Singh and Maurice Mars. Support Vector Machines to forecast changes in CD4 count of HIV-1 positive patients. Scientific Research and Essays; 2010, Vol.5 (17), pp.2384-2390.
11. Fayyad. Piatetsky-Shapiro and Smyth P. From data mining to knowledge discovery in databases. American Association for Artificial Intelligence; 1996.
12. Trybula, W.J. Data mining and knowledge discovery. Annual Review of Information Science and Technology; 1997, 32, 197-229.
13. Hand D, Mannila H and Smyth P. Principles of data mining. Massachusetts: Massachusetts Institute of Technology press; 2001.
14. Ian Witten H & Eibe F. Data mining: practical machine learning tools and techniques. Second Edition. San Francisco: Morgan Kaufmann Publishers, 2005.
15. Dzeroski S. Towards a General Framework for Data Mining. In: Dzeroski, S and Struyf, J (Eds.), Knowledge Discovery in Inductive Databases. LNCS 47474. Springer-Verlag; 2006.

16. Han, J. DMQL: A Data Mining Query Language for Relational Databases. In proceedings of DMKD-96 (SIGMOD-96 Workshop on KDD). Montreal: Canada; 1996.

17. Meo, R. An Extension to SQL for Mining Association Rules: Data Mining and Knowledge Discovery. Kluwer Academic Publishers; 1998, Vol(2), pp(195-224).

18. Imielinski T and Virmani A. MSQL: A Query Language for Database Mining. Data Mining and Knowledge Discovery. Kluwer Academic Publishers; 1999, Vol. 3, pp 373-408.

19. Sarawagi S. Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. *Data Mining and Knowledge Discovery*; 2000, Vol. 4, pp 89–125.

20. Botta M. Query Languages Supporting Descriptive Rule Mining: A Comparative Study. *Database Support for Data Mining Applications*; 2004, LNAI 2682, pp 24-51.

21. David L. and Dursen D. Advanced Data Mining Techniques. Berlin, Heidelberg: Springer-Verlag; 2008.

22. SAS Institute. *SAS Enterprise Miner – SEMMA* [Internet]. Available from: http://www.sas.com/technologies/analytics/datamining/miner/semma.html. [Retrieved on Jan 2013].

23. Santos M. & Azevedo C. Data Mining – Descoberta de Conhecimento em Bases de Dados. FCA Publisher; 2005.

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

25. Jiawei H and Micheline K. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers; 2006.

26. Ian H, Eibe F, and Mark A. Data Mining. Practical Machine Learning Tools and Techniques. Third Edition. 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA :Morgan Kaufmann Publishers is an imprint of Elsevier; 2011.

27. Ralf M and Markus R. Data Mining tools: Advanced Review. *WIREs Data Mining and Knowledge Discovery*. John Wiley & Sons, Inc.; 2011 00 1–13, DOI: 10.1002/widm.24

28. Informatics Research and Development Unit. Public Health Informatics & Technology Program Office. Office of Surveillance, Epidemiology, and Laboratory Services. US Centers for Disease Control and Prevention. 2010 Report: Open Source Data Mining Software Evaluation; 2010.

29. Kincade K. Data mining: digging for healthcare gold. *Insurance & Technology*; 1998 23(2), IM2-IM7.

30. Milley, A. Healthcare and data mining. *Health Management Technology*; 2000, 21(8), 44-47.

31. Christy, T. Analytical tools help health firms fight fraud. *Insurance & Technology*, 22(3), 22-26.

32. Biafore S. Predictive solutions bring more power to decision makers. *Health Management Technology*; 1999, 20(10), 12-14.

33. Silver, M. Sakata, T. Su, H.C. Herman, C. Dolins, S.B. & O'Shea, M.J. Case study: how to apply data mining techniques in a healthcare data warehouse. *Journal of Healthcare Information Management*; 2001, 15(2), 155-164.

34. Benko, A. & Wilson, B. Online decision support gives plans an edge. *Managed Healthcare Executive*; 2003, 13(5).

35. Gillespie, G. There's gold in them than databases. *Health Data Management*, 8(11), 40-52.

36. Kolar, H.R. Caring for healthcare. *Health Management Technology*; 2001, 22(4), 46-47.

37. Relles D. Ridgeway G & Carter G. Data mining and the implementation of a prospective payment system for inpatient rehabilitation. *Health Services & Outcomes Research Methodology*; 2002, 3(3-4), 247-266.

38. WHO. Antiretroviral Therapy [Internet].Available from: http://www.who.int/topics/antiretroviral_therapy/en/. [Retrieved on Dec 16, 2012.]

39. Factsheet on HIV and its treatment [Internet]. Available from: http://aidsinfo.nih.gov/guidelines. [Retrieved on Jan 28, 2013].

40. WHO. Facts about HIV/AIDS [Internet]. Available from: http://www.who.int/hiv/treatment/en/index.html. [Retrieved on: Jan 30, 2013].

41. WHO. Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach; 2010.

42. Darbyshire J. "Perspectives in drug therapy of HIV infection". Drugs 49 Suppl 1: 1–3; 1995. Discussion 38–40; 1997.

43. Ho D. "Time to hit HIV, early and hard". The New England Journal of Medicine; 1995, 333 (7): 450–451.

44. Harrington M, Carpenter C. "Hit HIV-1 hard, but only when necessary". The Lancet; 2000, 355 (9221): 2147.

45. Jain V, Deeks SG."When to start antiretroviral therapy". Current HIV/AIDS Rep; May 2010, 7 (2): 60–8.

46. Panel on Antiretroviral Guidelines for Adults and Adolescents. *Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents*. United States Department of Health and Human Services; 2009.

47. WHO. Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach– 2010 revision; 2010.

48. Elias L. HIV Status Predictive Modeling Using Data Mining Technology [Unpublished MSc Thesis]. Addis Ababa University: School of Information Science and School of Public Health; 2011.

49. Vararuk A, Petrounias I, Kodogiannis V. "Data mining techniques for HIV/AIDS data management in Thailand". *Journal of Enterprise Information Management*; 2008, 21(1), pp.52 – 70.

50. Rosma M, Sameem A, Basir A, Adeeba K, and Annapurni K. The Prediction of AIDS Survival: A Data Mining Approach. Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering.

51. Michael L, Edward A, and Thirumala G. Predictive Models for Maximum Recommended Therapeutic Dose of Antiretroviral Drugs: Computational and Mathematical Methods in Medicine. Hindawi Publishing Corporation; 2012, Article ID 469769, 9 pages doi:10.1155/2012/469769.

52. Melody Y, Kiang. A comparative assessment of classification methods. Information Systems Department, College of Business Administration, California State University, 1250 Bellflower Blvd., Long Beach, CA 90840, USA, doi:10.1016/S0167-9236(02)00110-0.

53. Thomas H. Mining Echocardiography data to predict heart disease. [Unpublished MSc Thesis]. Addis Ababa University: School of Information Science and School of Public Health; 2012.

54. Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations; 2009, Volume 11, Issue 1.

55. Lior R. and Oded M.Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing Co. Pte. Ltd; 2008.

56. Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995. Available from: http://www.springerlink.com/content/k238jx04hm87j80g/

57. Te-Ming Huang, Vojislav Kecman, IvicaKopriva. Kernel Based Algorithms for Mining Huge Data Sets. Springer-Verlag Berlin Heidelberg; 2006.

58. Ajith Abraham. Artificial Neural Networks. Oklahoma State University, Stillwater: USA.

59. Soumen C, Earl C, Eibe F, Ralf Hartmut G, Jaiwei H, Xia J, Micheline K, Sam S, Thomas P. Nadeau, Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J. Teorey, Ian H. Witten. Data Mining Know it All. Morgan Kaufmann Publishers; 2009.

60. Daniel T. Larose. Discovering Knowledge in Data: An introduction to Data Mining; A John Wiley & Sons, Inc. Publication, 2005.

61. WHO. Case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children; 2007, ISBN 978 92 4 159562 9.

62. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. Third Edition. USA: Two Crows Corporation; 2005.

63. UNICEF. Opportunities in crisis: Preventing HIV from early adolescence to young adulthood; June 2011.

64. Zhi-hua Z. Ensemble learning. National Key Laboratory for Novel Software Technology. China: Nanjing University.

65. Assareh A, Moradi H, Volkert G.A hybrid random subspace classifier fusion approach for protein mass spectra classification. In: Marchiori, E., Moore, J.H. (eds.) EvoBIO. LNCS, Springer, Heidelberg; 2008, vol. 4973, pp. 1–11.

66. Minale T. Application of Data Mining Techniques to Predict Urinary Fistula Surgical Repair Outcome [Unpublished MSc Thesis]. Addis Ababa University: School of Information Science and School of Public Health; 2012.

67. N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis Journal*, 6(5):429–450, November 2002

68. M. Kubat, R. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30:195–215, 1998.

69. Yanmin S, Mohammed S K, and Yang W. Boosting for Learning Imbalanced Class Distribution. Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 0-7695-2701-9/06; 2006.

70. Ajay T, Jamal A, Zubair S, Muddassar F. Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets. (Eds.): Pizzuti C, Ritchie M, and Giacobini M; EvoBIO: Springer-Verlag. Berlin: Heidelberg; 2009, LNCS 5483, pp. 128–139.

71. Zazzi M, Kaiser R, Sonnerborg A, Struck A, Altmann A, Prosperi M, Roosen-zvi M, Petroczi A, Petros Y, Schulter E, Boucher CA, Brun-vezinet F, Harigan PR, Morris L, Obermire M, Perno C-F, Phanuphak P, Pillay D, Shefer RW, Vandamme A-M, Van Laethem K, Wensing AMJ, Langauer T, Incardona F. Prediction of response to antiretroviral therapy by human experts and by the EuResist data-driven expert system. British HIV Association; 2010, DOI: 10.1111/j.1468-1293.2010.00871.x

72. FHAPCO.ART Scale-up in Ethiopia. Success and Challenges; 2009.

73. Nitesh V. Chawla. Data Mining for Imbalanced Dataset: An Overview. Department of Computer Science and Enginering. University of Notre Dame: USA.

Appendix A: Attribute Ranking for the Sixth Month CD4 Prediction

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: Last Modified2-weka.filters.unsupervised.attribute.Remove-R12-13

Instances: 7252

Attributes: 11

Age, Sex, MaritalStatus, EducationalStatus, FamilyPlanning, PregnancyStatus, FunctionalStatus, BaselineWHOStage, OriginalRegimen, BaselineCD4Count, CD4CountAfterSixMonths

Evaluation mode: evaluate on all training data

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

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 11 CD4CountAfterSixMonths):

Information Gain Ranking Filter

Ranked attributes:

0.21587	10 BaselineCD4Count	0.00973	8 BaselineWHOStage
0.01924	9 OriginalRegimen	0.00961	6 PregnancyStatus
0.01262	1 Age	0.0091	3 MaritalStatus
0.01195	5 FamilyPlanning	0.00871	2 Sex
0.01043	7 FunctionalStatus	0.00358	4 EducationalStatus

Selected attributes: 10, 9, 1, 5, 7, 8, 6, 3, 2, 4: 10

Appendix B: Attribute Ranking for the Twelfth Month CD4 Prediction

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: Last Modified2-weka.filters.unsupervised.attribute.Remove-R13

Instances: 7252

Attributes: 12

Age, Sex, MaritalStatus, EducationalStatus, FamilyPlanning, PregnancyStatus, FunctionalStatus, BaselineWHOStage, OriginalRegimen, BaselineCD4Count, CD4CountAfterSixMonths, CD4CountAftreTwelveMonths

Evaluation mode: evaluate on all training data

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

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 12 CD4CountAftreTwelveMonths):

Information Gain Ranking Filter

Ranked attributes:

0.4208	11 CD4CountAfterSixMonths	0.00853	6 PregnancyStatus
0.13315	10 BaselineCD4Count	0.00752	2 Sex
0.01163	5 FamilyPlanning	0.00542	7 FunctionalStatus
0.0107	1 Age	0.00302	8 BaselineWHOStage
0.01038	9 OriginalRegimen	0.00292	4 EducationalStatus
0.0096	3 MaritalStatus		

Selected attributes: 11, 10, 5, 1, 9, 3, 6, 2, 7, 8, 4: 11

Appendix C: Attribute Ranking for the Eighteenth Month CD4 Prediction

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: Last Modified2

Instances: 7252

Attributes: 13

Age, Sex, MaritalStatus, EducationalStatus, FamilyPlanning, PregnancyStatus, FunctionalStatus, BaselineWHOStage, OriginalRegimen, BaselineCD4Count, CD4CountAfterSixMonths, CD4CountAftreTwelveMonths, CD4CountAfterEighteenMonths

Evaluation mode: evaluate on all training data

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

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 CD4CountAfterEighteenMonths):

Information Gain Ranking Filter

Ranked attributes:

0.52416 12 CD4CountAftreTwelveMonths	0.01121 9 OriginalRegimen
0.3258 11 CD4CountAfterSixMonths	0.0101 6 PregnancyStatus
0.12897 10 BaselineCD4Count	0.00851 2 Sex
0.02215 5 FamilyPlanning	0.00732 4 EducationalStatus
0.01639 3 MaritalStatus	0.00454 7 FunctionalStatus
0.01397 1 Age	0.00309 8 BaselineWHOStage

Selected attributes: 12, 11, 10, 5, 3, 1, 9, 6, 2, 4, 7, 8: 12

Appendix D: Comparison of models generated using the Base Classifier and Boosting Algorithm

Model	Parameter	Classes			
		<200	**200 – 349**	**350 – 499**	**>=500**
J48 Unprunned	TPR	0.699	0.72	**0.535**	**0.417**
	FPR	0.123	0.26	0.099	0.033
	F-Measure	0.71	0.702	0.529	0.465
	AUC	**0.922**	**0.869**	**0.906**	**0.921**
J48 Unprunned and Boosted	TPR	**0.858**	**0.848**	**0.749**	**0.68**
	FPR	0.06	0.138	0.054	0.017
	F-Measure	0.863	0.838	0.742	0.725
	AUC	**0.956**	**0.933**	**0.948**	**0.958**

Table: Class-level Comparison between the Base and Boosted Unprunned J48 Classifiers

Model	Parameter	Classes			
		<200	**200 – 349**	**350 – 499**	**>=500**
PART Unprunned	TPR	0.646	0.717	**0.459**	**0.266**
	FPR	0.127	0.297	0.113	0.039
	F-Measure	0.671	0.684	0.454	0.310
	AUC	0.907	0.851	0.882	0.910
PART unprunned and boosted	TPR	0.859	0.848	**0.741**	**0.680**
	FPR	0.059	0.137	0.052	0.021
	F-Measure	0.864	0.838	0.742	0.707
	AUC	0.960	0.936	0.948	0.956

Table: Class-level Comparison between the Base and Boosted Unprunned PART Classifiers

From the above experimental results we can deduce that boosting has dramatically improved the TPR of the minority classes even though the improvements are also happened on the majority classes too.

Appendix E: Sample Weka output of the Selected Sixth month CD4 count Model

=== Run information ===

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART -- -U -M 2 -C 0.5 -Q 1

Relation: Last Modified2-weka.filters.unsupervised.attribute.Remove-R12-13

Instances: 7252

Attributes: 11

Age, Sex, MaritalStatus, EducationalStatus, FamilyPlanning, PregnancyStatus, FunctionalStatus, BaselineWHOStage, OriginalRegimen, BaselineCD4Count, CD4CountAfterSixMonths

Test mode: 10-fold cross-validation		Mean absolute error	0.1305
=== Classifier model (full training set) ===		Root mean squared error	0.2494
AdaBoostM1: Base classifiers and their weights:		Relative absolute error	38.6934 %
=== Stratified cross-validation ===		Root relative squared error	60.75 %
=== Summary ===		Coverage of cases (0.95 level)	97.6558 %
Correctly Classified Instances	5947 82.005 %	Mean rel. region size (0.95 level)	47.4111 %
Incorrectly Classified Instances	1305 17.995 %	Total Number of Instances	7252
Kappa statistic	0.7321		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.848	0.137	0.829	0.848	0.838	0.709	0.936	0.911	200 - 349
	0.741	0.052	0.742	0.741	0.742	0.69	0.948	0.832	350 - 499
	0.859	0.059	0.87	0.859	0.864	0.803	0.96	0.928	<200
	0.68	0.021	0.736	0.68	0.707	0.683	0.956	0.767	>=500
Weighted Avg:	0.82	0.089	0.82	0.82	0.82	0.733	0.947	0.892	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
2703	194	229	62	| a = 200 - 349
217	903	40	58	| b = 350 - 499
251	48	1948	21	| c = <200
90	72	23	393	| d = >=500

Appendix F: Sample Weka output of the Selected Twelfth Month CD4 Count Model

=== Run information ===

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART -- -U -M 2 -C 0.5 -Q 1

Relation: Last Modified2-weka.filters.unsupervised.attribute.Remove-R13

Instances: 7252

Attributes: 12

Age, Sex, MaritalStatus, EducationalStatus, FamilyPlanning, PregnancyStatus, FunctionalStatus, BaselineWHOStage, OriginalRegimen, BaselineCD4Count, CD4CountAfterSixMonths, CD4CountAftreTwelveMonths

Test mode: 10-fold cross-validation

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

AdaBoostM1: Base classifiers and their weights:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	6548	90.2923%	Mean rel. region size (0.95 level)	28.0819 %
Incorrectly Classified Instances	704	9.7077%	Total Number of Instances	7252
Kappa statistic	0.8568			

Mean absolute error	0.0482
Root mean squared error	0.1834
Relative absolute error	14.1996 %
Root relative squared error	44.5199 %
Coverage of cases (0.95 level)	96.5389 %

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.911	0.017	0.915	0.911	0.913	0.895	0.984	0.946	<200
	0.868	0.042	0.874	0.868	0.871	0.828	0.971	0.922	350 - 499
	0.921	0.075	0.914	0.921	0.918	0.846	0.969	0.959	200 - 349
	0.895	0.012	0.9	0.895	0.898	0.885	0.982	0.924	>=500
Weighted Avg:	0.903	0.05	0.903	0.903	0.903	0.854	0.973	0.944	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
1135	15	94	2	a = <200
17	1592	171	54	b = 350 - 499
86	156	3098	24	c = 200 - 349
2	58	25	723	d = >=500

Appendix G: Sample Weka output of the Selected Eighteenth Month CD4 Count Model

=== Run information ===

Scheme: weka.classifiers.rules.PART -M 2 -C 0.5 -Q 1

Relation: Last Modified2

Instances: 7252

Attributes: 13

Age, Sex, MaritalStatus, EducationalStatus, FamilyPlanning, PregnancyStatus, FunctionalStatus, BaselineWHOStage, OriginalRegimen, BaselineCD4Count, CD4CountAfterSixMonths, CD4CountAftreTwelveMonths, CD4CountAfterEighteenMonths

Test mode: 10-fold cross-validation			Mean absolute error	0.1431
=== Classifier model (full training set) ===			Root mean squared error	0.3305
=== Stratified cross-validation ===			Relative absolute error	41.2275 %
=== Summary ===			Root relative squared error	79.3364 %
Correctly Classified Instances	5085	70.1186 %	Coverage of cases (0.95 level)	95.16 %
Incorrectly Classified Instances	2167	29.8814 %	Mean rel. region size (0.95 level)	37.1208 %
Kappa statistic	0.5687		Total Number of Instances	7252

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.622	0.045	0.64	0.622	0.631	0.584	0.948	0.751	<200
	0.697	0.165	0.699	0.697	0.698	0.532	0.892	0.808	350 - 499
	0.727	0.176	0.716	0.727	0.721	0.549	0.899	0.835	200 - 349
	0.706	0.051	0.714	0.706	0.71	0.658	0.949	0.811	>=500
Weighted Avg:	0.701	0.138	0.701	0.701	0.701	0.564	0.91	0.812	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
514	67	237	8	| a = <200
56	1792	491	233	| b = 350 - 499
217	463	2001	71	| c = 200 - 349
16	242	66	778	| d = >=500