ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH: HEALTH INFORMATICS PROGRAM

CONSTRUCTING A PREDICTIVE MODEL FOR DETERMINING CD4 STATUS OF PATIENTS FOLLOWING ART: THE CASE OF JIMMA AND BONGA HOSPITALS

BEHAILU G/MARIAM

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH: HEALTH INFORMATICS PROGRAM

CONSTRUCTING A PREDICTIVE MODEL FOR DETERMINING CD4 STATUS OF PATIENTS FOLLOWING ART: THE CASE OF JIMMA AND BONGA HOSPITALS

A thesis submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

BEHAILU G/MARIAM (BEd)

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH: HEALTH INFORMATICS PROGRAM

CONSTRUCTING A PREDICTIVE MODEL FOR DETERMINING CD4 STATUS OF PATIENTS FOLLOWING ART: THE CASE OF JIMMA AND BONGA HOSPITALS

A thesis submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

Members of the Examination Board

Name Signature

1. Dr.Dereje Teferi (Advisor)_____
2. Dr.Getinet Mitike (Advisor)_____
3. Dr.Milion Meshesha (Examiner)_____
4. Dr.Jemal Haidar (Examiner)_____

DEDICATION

This paper is dedicated to my mother and father, who have always dreamed to see my successes.

DECLARATION

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

_____

Behailu G/mariam

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

_____

Dereje Teferi (PhD)

and

_____

Getinet Mitike (PhD)

ACKNOWLEDGEMENT

First of all I would like to thank my advisors Dr. Dereje Teferi and Dr. Getinet Mitike for their constructive, valuable and supportive ideas. My advisors patience to improve my work always encouraged me to complete the research with high interest.

I would like to forward my thanks to the Jimma and Bonga hospital managers, who facilitate for me to obtain the necessary data. My special thanks go to Ato Tewekel Asrat data clerk of the Bonga hospital at the moment. I also like to thank W/ro Genet, Jimma hospital data clerk her support was unforgettable.

I would also like to thank members of the School of information science. My special thanks go to Ato Mahider Alemayehu, who helped me a lot when I join the department. W/rt Meseret Ayano's coordination is smart and I always remember her respect for people. W/rt Misraq is also very helpful and communicable person.

My pleasure and deepest gratitude goes to my families. I always remember the true love my sister Mekides G/mariam gives to me all the time. I would like to forward my respect to Ato. Frehiwot Getahun, I always remember his logistics and idea support for my education.

Finally my adore Konjit H/Michael took the best of all respect, gratitude, thanks and love in my life. With out her support every thing in my research work was unthinkable. Konjit gave me mental peace, confidence and love in doing the research work.

ABSTRACT

Background: Many of the reports on HIV/AIDS shows that the number of ART registered patients are increasing from time to time. However those reports show that the increasing of patient's number, they did not try to make prediction of attributes based on the given attributes more than statistical explanation. This study concerned to use data mining technique on ART data base. The study data was taken from two hospitals of the south west of Ethiopia namely Jimma and Bonga hospitals.

Objective: The main objective of the study is to integrate the applicability of data mining techniques on predicting CD4 status of patients following ART in Jimma, and Bonga Hospitals. The main goal of this research is to find the pattern of attributes of the patient in order to build predictive model using data mining techniques.

Methodology: The study followed the CRISP-DM data mining methodology, which has six phases called: business understanding, data understanding, data preparation, model building, evaluation and deployment. The study used classification to predict the status of CD4 of patients following ART. J48 is a technique used for building classification and PART is used to compare the result of J48.

Findings: The best performance achieved by J48 decision tree algorithm is a generalized decision tree with pruning with reduced attributes. The model classifies instances correctly 88.79% and incorrectly classifies 11.21%. The weighted average precision of the model is 0.88 with recall of 0.89 and ROC area of 0.85. The model has 760 numbers of leaves and 916 tree size. The time taken to build the model is 0.05 seconds. The analysis of this model shows that the model is quit efficient to predict CD4 status of patients following ART.

Conclusion: Classification done using J48 decision tree is the best model than PART rule induction algorithm. J48 algorithm is effective to predict the CD4 status of patients following ART. From the model built it is fund that attributes: Eligible reason, ART status, ART start year, OAweight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status, Past ARV are the most determining factors of CD4 status.

CHAPTER ONE

INTRODUCTION

1.1 Background

1.1.1 HIV/AIDS profile in Ethiopia

The first case of HIV in Ethiopia was reported in 1984. Since then, HIV/AIDS has become a major public health concern in the country, leading the Government of Ethiopia to declare a public health emergency in 2002. In 2007, the estimated adult HIV/AIDS prevalence in Ethiopia was 2.1 percent. Although the epidemic is currently stable, HIV/AIDS remains a major development challenge for Ethiopia. Poverty, food shortages, and other socio-economic factors amplify the impact of the epidemic. According to the most recent data from the Joint United Nations Program on HIV/AIDS (UNAIDS), approximately 980,000 Ethiopians were living with HIV/AIDS in 2007, and 67,000 individuals have died as a result of infection with the virus. National projections estimate approximately 1.1 million Ethiopians are living with HIV and prevalence increased slightly to 2.3 percent by 2009 (USAID, 2010).

In the last few years, the Government of Ethiopia has increased efforts to accelerate progress toward universal access to HIV prevention, treatment, care, and support, emphasizing involvement of local stakeholders and decentralization. A two-year Millennium AIDS Campaign to improve access to treatment was launched in 2006, which included efforts to decentralize the HIV/AIDS response; broad-based communications; coordinated planning; specific performance targets; and improved integration of HIV/AIDS treatment into health care settings. The Government trained 32,000 health extension workers to promote community health initiatives, aid households, and deliver ART as part of the decentralization. According to Ethiopia's Ministry of Health, the Campaign reached approximately 1 million people, and the number of ART recipients increased more than 14-fold from 2005 to 2007 (USAID, 2010).

1.1.2 The Impact of ART in Ethiopia

Antiretroviral Therapy (ART) is treatment for AIDS that helps the body's immune system recover from the damage caused by infection with HIV. Although ART cannot cure AIDS, persons on ART will begin to feel better, eat more, and put on weight. Their bodies will recover the ability to fight infections. As persons on ART treatment become well, they can care for their children and return to household activities and productive life, which benefits the household and national economies. They recover their sense of hope for the future and can become powerful advocates for prevention and mitigation of HIV in their families and communities. They may remain well for many years, but must continue to take Antiretroviral (ARVs) for the rest of their lives. Thus, ART is an important component of the global response to AIDS (Path finder International, 2007).

The technical document for the sixth report of Aids in Ethiopia reported that there is a need to ART 351,001 in the year 2010. The HIV prevalence is expected to increase as the number of people taking ART increase or survive longer. If universal access to ART is achieved, the national HIV prevalence will slightly increase from 2.8% without ART to 3.1% with ART in 2010.

The same report stated the potential effect of ART on AIDS death, from 2000-2010 assuming that a successful implementation of the MOH's ART rollout plan. The number of AIDS deaths will start to decline from 2005 onwards. By the year 2010, there will be 41% fewer AIDS death compared to a projection without an ART program (MOH/HAPCO, 2006).

1.1.3 When to start antiretroviral therapy in adults and adolescents

WHO guideline: Antiretroviral therapy for HIV infection in adults and adolescents:

Recommendations for a public health approach (WHO, 2006) defines when to start ART.

> *In resource limited settings the decision to initiate ART in adults and adolescents relies on clinical and immunological assessment. In order to facilitate the rapid scale-up of ART programs with a view to achieving universal access to this therapy, WHO emphasizes the importance of using clinical parameters in deciding when to*

initiate it. However, it is recognized that the value of clinical staging in deciding when to initiate and monitor ART is improved by additional information on baseline and subsequent (longitudinal) CD4 cell counts. While WHO continues to advocate wider availability of affordable point-of-care CD4 cell count testing, the lack of a CD4 count should not delay the initiation of ART if the patient in question is clinically eligible. WHO encourages national programs to increase access to CD4 measurement technologies. The process of initiating ART involves assessing patient readiness to commence therapy and an understanding of its implications (lifelong therapy, adherence, toxicities).

1.1.4 What are CD4 Cells?

CD4 cells are a type of lymphocyte (white blood cell). They are an important part of the immune system. CD4 cells are sometimes called T-cells. There are two main types' of T-cells. T-4 cells, also called CD4, are "helper" cells. They lead the attack against infections.HIV most often infects CD4 cells. The genetic code of the virus becomes part of the cells. When CD4 cells multiply to fight an infection, they make more copies of HIV. When someone is infected with HIV but has not started treatment, the number of CD4 cells they have goes down. This is a sign that the immune system is being weakened. The lower the CD4 cell count, the more likely the person will get sick. There are millions of different families of CD4 cells. Each family is designed to fight a specific type of germ.

When HIV reduces the number of CD4 cells, some of these families can be wiped out. The person can lose the ability to fight off the particular germs those families were designed for. If this happens, the person might develop an opportunistic infection (Fact sheet, 2011).

1.1.5WHO clinical stages

The clinical stages of HIV stage adopted from the antiretroviral therapy for HIV infection in adults and adolescents: Recommendations for a public health approach 2006(WHO, 2006) classifies the stages of HIV based on the clinical stages as follows:

CLINICAL STAGE 1

- Asymptomatic.
- Persistent generalized lymphadenopathy.

CLINICAL STAGE 2

- Unexplained a moderate weight loss (under 10% of presumed or measured body weight).
- Recurrent upper respiratory tract infections (sinusitis, tonsillitis, otitis media, pharyngitis).
- Herpes zoster.
- Angular cheilitis.
- Recurrent oral ulceration.
- Papular pruritic eruptions.
- Seborrhoeic dermatitis.
- Fungal nail infection.

CLINICAL STAGE 3

- Unexplained severe weight loss (over 10% of presumed or measured body weight).
- Unexplained chronic diarrhoea for longer than one month.
- Unexplained persistent fever (intermittent or constant for longer than one month).
- Persistent oral candidiasis.
- Oral hairy leukoplakia.
- Pulmonary tuberculosis (current).
- Severe bacterial infections (e.g. pneumonia, empyema, pyomyositis, bone or joint infection, meningitis, bacteraemia, severe pelvic inflamatory disease).
- Acute necrotizing ulcerative stomatitis, gingivitis or periodontitis.
- Unexplaineda 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 pneumonia.
- Recurrent bacterial pneumonia.
- Chronic herpes simplex infection (orolabial, genital or anorectal of more than one month's duration or visceral at any site).
- Oesophageal candidiasis (or candidiasis of trachea, bronchi or lungs).
- Extrapulmonary tuberculosis.
- Kaposi sarcoma.
- Cytomegalovirus infection (retinitis or infection of other organs).
- Central nervous system toxoplasmosis.
- HIV encephalopathy.
- Extrapulmonary cryptococcosis including meningitis.
- Disseminated non-tuberculous mycobacteria infection.
- Progressive multifocal leukoencephalopathy.
- Chronic cryptosporidiosis.
- Chronic isosporiasis.
- Disseminated mycosis (coccidiomycosis or histoplasmosis).
- Recurrent septicaemia (including non-typhoidal *Salmonella*).
- Lymphoma (cerebral or B cell non-Hodgkin).
- Invasive cervical carcinoma.
- Atypical disseminated leishmaniasis.
- Symptomatic HIV-associated nephropathy or HIV-associated cardiomyopathy.

1.1.6 ART and data mining

Many of the reports on HIV/AIDS shows that the number of ART registered patients are increasing from time to time. However those reports show that the increasing of patient's number, they did not try to make prediction of attributes based on the given attributes more than statistical explanation. This large number of registry of patients on ART includes a lot of attributes that one patient's record holds.

The attributes are very much useful in order to explain one or more dependent variable. The prediction of dependent variable from the independents need different technique called data mining technique.

Two Crows Corporation, (2005) define data mining as: *a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.*

1.2 Statement of the problem

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. The ART-LINC collaboration (18 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. 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 (WHO-HIV/AIDS Program, 2006).

The CD4 cell count remains the strongest predictor of HIV related complications, even after the initiation of therapy. The baseline pretreatment value is informative: lower CD4 counts are associated with smaller and slower improvements in counts. However, precise thresholds that define treatment failure in patients starting at various CD4 levels are not yet established. As a general rule, new and progressive severe immunodeficiency as demonstrated by declining longitudinal CD4 cell counts should trigger a switch in therapy (Gadelha, 2002).

Patients starting with low CD4 count may demonstrate slow recovery, but persistent levels below 100 cells/mm^3 represents significant risk for HIV disease progression. Caveats 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. As a general principle, inter current infections should be managed, time should be allowed for recovery and the CD4 cell count should

be measured before ART is switched. If resources permit, a second CD4 cell count should be obtained to confirm immunological failure. Reasonable working definitions of immunological failure are: (Mellors, 1997). (1) CD4 count below 100 cells/mm3 after six months of therapy; (2) a return to, or a fall below, the pre therapy CD4 baseline after six months of therapy; or (3) a 50% decline from the on treatment peak CD4 value (if known).

The CD4 cell count can also be used to determine when not to switch therapy, e.g. in a patient with a new clinical stage 3 event for whom switching is being considered or in a patient who is asymptomatic and under routine framework. In general, switching should not be recommended if the CD4 cell count is above 200 cells/mm^3 (Mellors, 1997).

As we have seen in the above paragraphs CD4 cell count has many advantages on the patients ART follow up. Since it is difficult to measure the CD4 cell count every time, and will create unnecessary anxiety on patients it is necessary to prepare a CD4 cell count status prediction model. The main concern of this research is to predict those patients who will develop low CD4 count status and high CD4 count status after three months. Thus the goal of this research is to apply data mining techniques and predict CD4 status of patients using the data set of ART database.

The research outcome will have a contribution to identify the important patterns of ART data set in order to make decision and intervention to the result. In doing so this research will answer the following questions:

- What are the main attributes affecting CD4 status?
- Which data mining technique is suitable for predicting CD4 status of patients?

1.3 Objective of the study

1.3.1 General objective

The main objective of the study is to apply data mining techniques for predicting CD4 status of patients on ART in Jimma, and Bonga Hospitals.

1.3.2 Specific objectives

The specific objectives are to:

- Identify the important attributes affecting CD4 status.
- Develop patterns and classify patients in to increasing and decreasing CD4 status.
- Compare the result of Experiments using different techniques.

1.4 Research Modeling

1.4.1 Study design/Modeling Data Mining

In this research CRIPS-DM (Cross Industry Standard Process for Data Mining) modeling is used for Business understanding, data understanding, data preparation and pre processing, model building, evaluation and deployment steps. (CRIPS-DM is also known as six phases which that shows the phases of a data mining process.)

The sequence of the phases is not strict. Moving back and forth between different phases is always required. It depends on the outcome of each phase which phase, or which particular task of a phase, has to be performed next. The six phases are shown in fig 1.1 The arrows indicate the most important and frequent dependencies between phases.

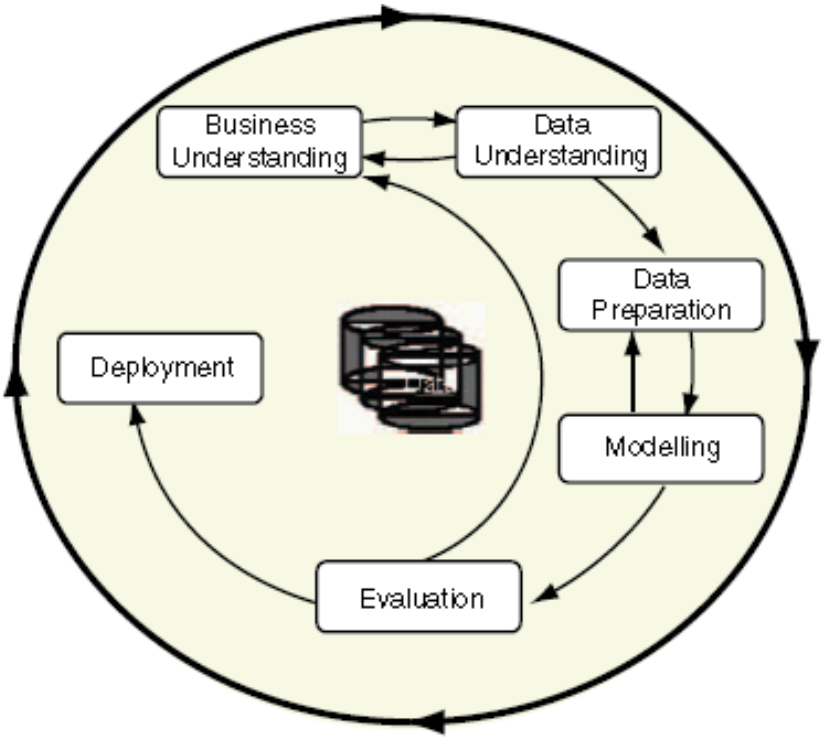

Fig 1.1 Phases of the CRISP-DM model Source: Larose et al, (2005)

1.4.2.1 Understanding Business/Problem

The main objective of the health sectors is to improve the health status of the patients. Especially the ART service delivery health facilities including Bonga and Jimma hospitals targeted to improve the CD4 status of patients. Based on the main objective the researcher is able to formulate the following questions:

- What are the main attributes affecting CD4 status?
- Which data mining technique is suitable for predicting CD4 status of patients?

The plan is to apply appropriate data mining techniques to try to uncover these and other possible patterns.

1.4.2.2 Data collection and understanding

In this phase data collected from the original sources of ART data bases of Hospitals of Jimma and Bonga towns. Use exploratory data analysis to familiarize with the data and discover initial insights. Identify the data type of attributes. Evaluate the quality of the data.

1.4.2.3 Data preparation and pre-processing

In this phase data cleaned such as how to handle missing values, outliers and noisy data and data will be converted to the necessary format. The missing values are handled by replacing it to modal values for nominal attributes and mean values for numeric data type. Transformation of weight attribute is done by converting the numbers to labeled class and CD4 attribute changed to nominal values based on the guide line.

1.4.2.4 Model Building

This phase is the step to select techniques of data mining and applying different settings for the selected techniques. In doing so, selection and applying appropriate modeling techniques is done. Calibrate model settings to optimize results done by changing the parameter of J48 in to “Binary” and “General”.

Since modeling needs to consider different considerations of the application of the technique, classification is one of the techniques used for prediction purpose of the study. In order to develop a predictive and classification model, decision tree classification data mining technique is done. For the purpose of rule generation and comparison of results PART rule induction algorithm is performed.

1.4.2.5 Analyze the result

In this phase analyzing the result has been done. Analyzing the result of J48 is done by comparing the precision and accuracy of the experiments done. Out of the total experiments one selected. PART rule induction algorithm is performed by the same settings of the selected J48 algorithm. Comparison of results of J48 and PART is done by comparing the precisions and accuracies of the two experiments. Finally it is decided to use the 8^{th} experiment for building the model.

1.4.2.6 Deployment

In this phase deployment of the selected model setting is done in order to produce rules and to extract rules from the tree generated. The extracted rules are taken for pattern generating.

1.4.3 Tools

The following tools have been used in the research work:

- **Ms-Excel:** used for the purpose of data preparation and pre-processing.
- **Sys tool recovery:** used for the purpose of coping SQL records into Ms-Excel.
- **Weka:** used for the purpose of building models, evaluation and analysis.

1.4.4 Ethical consideration

Since the study is going on secondary data the researcher does not have personal contact to the patients and no need to ask consents. The research is governed for the following ethical considerations:

- The researcher does nothing with personal identifiers.
- The research is purely dedicated to academic purpose.
- Ethical clearance obtained from ethical clearance committee of school of Public health.
- The research is purely for public benefits.
- The research does not have harm on anybody in any ways.

1.5 Scope of the study

The scope of this research is only limited to predicting CD4 status of ART following patients of Jimma and Bonga Hospitals. Moreover the study is limited to develop a predictive model.

1.6 Limitations of the study

In doing the study the researcher faces different challenges, by leaving the silly challenges the most acknowledged are the following.

- Lack of standard between ART databases of Jimma and Bonga Hospitals leads to have different format. Creating one format for the two hospitals increase the complexity of research work.
- As part of data mining work handling noisy and missing values are very difficult. The difficulty of the work is raised from the less attention of record keepers.
- Shortage of reference and similar work on data mining and ART records limit the researcher to focus on minimum references.
- Obtaining ethical clearance and schedule of the department for the research work are not following the same path, this situation creates wastage of time.

1.7 Significance of the study

Data mining technology helps in to find patterns in data that are valid, novel, useful, and understandable. The pattern which has discovered can help the health care service provider decision makers in order to give knowledge full decisions. Predicting CD4 status from the ART dataset of patients following ART is very important in order to reduce the complication of patients.

The main goal of this research is to find the pattern of attributes of the patient in order to build predictive model using data mining techniques. The built model can help to predict which patient's CD4 will decrease or increase after a known period of time. Based on this predicted value the patient and the concerned health service delivery body can control the situation before it happen using different methods.

1.8 Organization of the thesis

This thesis is divided in to six chapters and the chapters do have main topics and sub topics. The organization includes the main topics of the chapters only.

Chapter one includes: Background of the study, statement of the problem, Objective of the study, Research Modeling (methodology), Scope of the study, limitation of the study and Significance of the study.

Chapter two includes: Literatures reviews on Concept of Data mining, The Knowledge Discovery Process, Tasks of Data Mining, Data Mining on Electronic Health Records, and Related works.

Chapter three mainly working on Data preprocessing and model selection: Data preprocessing, Data Description, Statistical description of attributes, Data cleaning, Model implementation.

Chapter four works on Model building: Attribute ranking, Classification model building, using different scenarios for both J48 decision tree algorithm and PART rule induction algorithms.

Chapter five discussed on Experiment and analysis of classification model: Experiment description, J48 algorithms model building including the results of the experiments and result analysis of experiments, Generating rules from decision tree this topic includes the extracted rules from the selected model, Rule generating using PART rule induction Algorithm this topic includes the results of PART and extracted rules and the comparison of J48 decision tree and PART algorithms

Chapter six deals with forwarding: Summary of the study, Conclusions and Recommendations based on findings from the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Concept of Data mining

Database technology has been used with great success in traditional business data processing. There is an increasing desire to use this technology in new application domains. One such application domain that is likely to acquire considerable significance in the near future is database mining. An increasing number of organizations are creating ultra large databases (measured in gigabytes and even terabytes) of business data, such as consumer data, transaction histories, sales records, etc.; such data forms a potential gold mine of valuable business information. Data mining is a relatively new and promising technology. It can be defined as the process of discovering meaningful new correlation, patterns, and trends by digging into (mining) large amounts of data stored in warehouse, using statistical, machine learning, artificial intelligence (AI), and data visualization techniques. Industries that are already taking advantage of data mining include medical, manufacturing, aerospace, chemical, etc. Knowledgeable observers generally agree that in-depth decision support requires new technology. This new technology should enable the discovery of trends and predictive patterns in data, the creation and testing of hypothesis, and generation of insight-provoking visualizations. Data mining helps the end users to extract useful information from large databases. These large databases are present in data warehouses, i.e., “Data Mountain,” which are presented to data mining tools. In short data warehousing allows one to build the data mountain (Sumathi et.al, 2006).

Data mining is the nontrivial extraction of implicit, previously unknown and potentially useful information from the data mountain. This data mining is not specific to any industry – it requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data. Data mining is also referred to as knowledge discovery in databases (KDD) (Sumathi et.al, 2006).

2.1.1 Types of information

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files (Zaïane, 1999).

Business transactions: Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house wares and assets. Large department stores, for example, thanks to the widespread use of bar codes, store millions of transactions daily representing often terabytes of data. Storage space is not the major problem, as the price of hard disks is continuously dropping, but the effective use of the data in a reasonable time frame for competitive decision making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world.

Scientific data: Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, or in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated.

Medical and personal data: From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele. Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared. When correlated with other data, this information can shed light on customer behavior and the like.

Surveillance video and pictures: With the amazing collapse of video camera prices, video cameras are becoming ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a tendency today to store the tapes and even digitize them for future use and analysis.

Satellite sensing: There are a countless number of satellites around the globe: some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, receives more data every second than what all NASA researchers and engineers can cope many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.

Games: Our society is collecting a tremendous amount of data and statistics about games, players and athletes. From hockey scores, basketball passes and car-racing lapses, to swimming times, boxer's pushes and chess positions, and all the data are stored. Commentators and journalists are using this information for reporting, but trainers and athletes would want to exploit this data to improve performance and better understand opponents.

Digital media: The proliferation of cheap scanners, desktop video cameras and digital cameras is one of the causes of the explosion in digital media repositories. In addition, many radio stations, television channels and film studios are digitizing their audio and video collections to improve the management of their multimedia assets. Associations such as the NHL and the NBA have already started converting their huge game collection into digital forms.

CAD and Software engineering data: There are a multitude of Computer Assisted Design (CAD) systems for architects to design buildings or engineers to conceive system components or circuits. These systems are generating a tremendous amount of data. Moreover, software engineering is a source of considerable similar data with code, function libraries, objects, etc., which need powerful tools for management and maintenance.

Virtual Worlds: There are many applications making use of three-dimensional virtual spaces. These spaces and the objects they contain are described with special languages such as VRML. Ideally, these virtual spaces are described in such a way that they can share objects and places. There is a remarkable amount of virtual reality object and space repositories available. Management of these repositories as well as content-based search and retrieval from these repositories are still research issues, while the size of the collections continues to grow.

Text reports and memos (e-mail messages): Most of the communications within and between companies or research organizations or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.

The World Wide Web repositories: Since the inception of the World Wide Web in 1993, documents of all sorts of formats, content and description have been collected and inter-connected with hyperlinks making it the largest repository of data ever built. Despite its dynamic and unstructured nature, its heterogeneous characteristic, and its very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers. Many believe that the World Wide Web will become the compilation of human knowledge (Zaïane, 1999).

2.1.2 What is Data Mining?

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting meaningful knowledge. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that were traditionally too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable

one. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides (Sumathi.et.al, 2006).

2.1.3 The Evolution and future of Data Mining

Sumathi and Sivanandam (2006) discuss on the evolution of data mining and future of data mining as follows:

> *Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysts. Traditional query and report tools have been used to describe and extract what is in a database. The user forms a hypothesis about a relationship and verifies it or discounts it with a series of queries against the data. For example, an analyst might hypothesize that people with low income and high debt are bad credit risks and query the database to verify or disprove this assumption. Data mining can be used to generate a hypothesis. For example, an analyst might use a neural network to discover a pattern that analysts did not think to try for example, that people over 30 years with low incomes and high debt but who own their own homes and have children is good credit risks. In the short term, the results of data mining will be in profitable, if mundane, business-related areas. Micromarketing campaigns will explore new niches. Advertising will target potential customers with new precision. In the medium term, data mining may be as common and easy to use as e-mail. We may use these tools to find the best airfare to New York, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers. The long-term prospects are truly exciting. Imagine intelligent agents turned loose on medical research data or on subatomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe.*

2.2 The Knowledge Discovery Process

Data mining and knowledge discovery projects involve a complex process of identifying, understanding, and transforming data modeling and evaluation efforts. Naturally, any models that are useful would need to be deployed before making an impact in practice.

There are several data mining process frameworks that seek to organize these activities (Fayyad et al., 1996).

A widely cited industry framework is the Cross-Industry Standard Process for Data Mining (CRISP-DM) model Larose et al, (2005). CRISP-DM offers a general model for data/text mining projects, highlighting the key tasks involved. According to the CRISP-DM framework, the life cycle of a knowledge discovery project consists of six phases, but the sequence of the phases is not strictly applied. Moving back and forth between different phases is always required. The process is iterative because the choice of subsequent phases often depends on the outcome of preceding phases.

The life cycle begins with business understanding to ground the overall aims of the project, and then moves to data understanding to identify potential inputs and outputs, data quality issues, and potential privacy or security concerns. The third phase, data preparation, involves the extraction of relevant data for a particular modeling effort, data quality assurance, and any transformations required for specific modeling techniques (Two crows corporation, 1999). Typically, the data preparation tasks account for the majority of effort about 60% in a data mining project. The fourth phase, data modeling, is the central focus of any knowledge discovery effort and consists of the construction of models based on a variety of techniques, with evaluations (the fifth phase) conducted for all modeling techniques. The final step is deployment so that useful models can be embedded in information systems to support decision-making activities (Sumathi.et.al, 2006).

The steps of Knowledge Discovery is shown below in fig 2.1

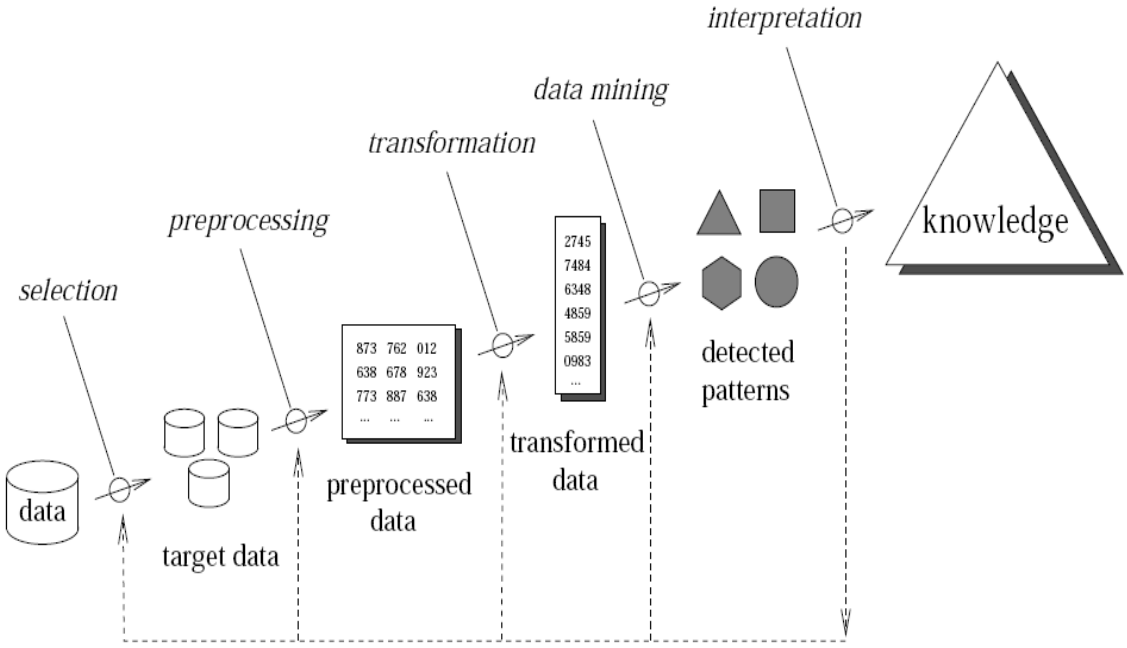

Fig 2.1 Steps of Knowledge Discovery

2.3 Tasks of Data mining

Data mining tasks mainly includes Exploratory Data Analysis (EDA), Descriptive Modeling, Predictive Modeling, Discovering Patterns and Rules and Retrieval by Content. Exploratory Data Analysis (EDA) helps simply to explore the data without any clear ideas of what we are looking for. Typically, EDA techniques are interactive and visual, and there are many effective graphical display methods for relatively small, low-dimensional data sets (David et al., 2001).

The goal of a descriptive model is describe all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (density estimation), partitioning of the p-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling) (David et al., 2001).

Predictive Modeling: Classification and Regression. The aim here is to build a model that will permit the value of one variable to be predicted from the known values of other variables. In classification, the variable being predicted is categorical, while in regression the variable is quantitative. The term "prediction" is used here in a general sense, and no notion of a time continuum is implied. So, for example, while we might want to predict the value of the stock market at some future date, or which horse will win a race, we might also want to determine the diagnosis of a patient, or the degree of brittleness of a weld (David et al., 2001).

One of the tasks of DM is Discovering Patterns and Rules. The Patterns and Rules task for data mining is the job of finding which attributes “go together.” Most prevalent in the business world, where it is known as affinity analysis or market basket analysis, the task of association seeks to uncover rules for quantifying the relationship between two or more attributes. Association rules are of the form “If antecedent, then consequent,” together with a measure of the support and confidence associated with the rule (Larose.et.al, 2006).

Retrieval by Content task is most commonly used for text and image data sets. For text, the pattern may be a set of keywords, and the user may wish to find relevant documents within a large set of possibly relevant documents (e.g., Web pages). For images, the user may have a sample image, a sketch of an image, or a description of an image, and wish to find similar images from a large set of images. In both cases the definition of similarity is critical, but so are the details of the search strategy (David. et.al, 2001).

2.3.1 Classification

Classification is a process of learning a function that maps a data item into one of several predefined classes (Larose.et.al, 2006). Every classification based on inductive-learning algorithms is given as input a set of samples that consist of vectors of attribute values (also called feature vectors) and a corresponding class. The goal of learning is to create a

classification model, known as a classifier, which will predict, with the values of its available input attributes, the class for some entity (a given sample). In other words, as defined by Larose.et.al (2006) classification is the process of assigning a discrete label value (class) to an unlabeled record, and a classifier is a model (a result of classification) that predicts one attribute class of a sample when the other attributes are given. In doing so, samples are divided into predefined groups. For example, a simple classification might group customer billing records into two specific classes: those who pay their bills within thirty days and those who takes longer than thirty days to pay.

Different classification methodologies are applied today in almost every discipline where the task of classification, because of the large amount of data, requires automation of the process. Examples of classification methods used as a part of data-mining applications include classifying trends in health records, financial market and identifying objects in large image databases (Kantardzic.M.2003).

2.4 Data Mining on Electronic Health Records

Data mining can be described as the inductive approach to uncovering patterns in data. In the particular case of data mining on EHRs, the process requires the application of machine learning techniques, or more precisely, knowledge discovery in databases. The use of data mining techniques could allow the classification of patients by risk category, or could help predict the length of hospitalization for a specific patient after a surgical procedure. The knowledge discovery process is as a multistep life cycle that begins with problem and data understanding, explicitly highlights the large effort necessary for data preparation, and then proceeds to modeling and evaluation. The final phase is the deployment of predictive models into existing systems. Data mining techniques lie at the intersection of machine learning, artificial intelligence, database systems, and statistics. The combination of these disciplines allows the extraction of hidden patterns from the underlying data (Kantardzic, 2003).

Machine learning algorithms provide inductive, data-driven approaches to a wide variety of tasks, such as classification and prediction. Classification tasks try to group individual

data entries into a known set of categories. Predictive tasks look into the future, applying a model to new data and making a qualitative classification or even a quantitative estimation (Kantardzic, 2003).

In addition to machine learning, databases and data warehouses also play a key role in the data mining process. The original roles of database systems data collection and database creation shifted toward improved data management, incrementing their functionality with sophisticated transaction processing techniques and technologies that allow the knowledge discovery process. In the healthcare industry, data warehousing tools have been successfully applied for planning and decision support in both the private and public sectors (Sumathi, et.al.2006).

Modern data mining incorporates many statistical techniques that aid in the tasks of feature selection and extraction. Statistical techniques such as correlation analysis, histograms, and principal component analysis and tests are incorporated for the processes of data visualization, attribute selection, and outlier investigation, and also to correct models so that they do not "over fit" the data; and finally, statistical techniques are used to evaluate data mining models and to express their significance (Hristidis., 2010).

2.5 Related works

Dr. Toshio Makie, MD, PhD investigates a research with a title "Estimating $CD4^+$ Cell Counts of an Individual using Population Historical Data". Makie (2008) discusses the background and the objective as follows:

"WHO guidelines in 2010 no longer recommend the total lymphocyte count (TLC), instead of the $CD4^+$ cell (CD4) count, in deciding on the initiation of anti-retroviral treatment. This is very problematic because of the cost and resources involved in getting CD4 counts. To solve this problem, we developed an efficient and reliable method for predicting CD4 counts based on a previous CD4 count."

The subject and methods used in the work is as follows:

"Of 832 HIV-infected patients registered at Osaka National Hospital between June 1998 and March 2006, 151 were immediately treated upon arrival, were already under treatment or in withdrawal. The rest, or the 681 antiretroviral-naïve, constitute our study sample. These patients underwent a total of 4,370 routine examinations that included determining CD4 counts and basic blood parameters such as hemoglobin (*Hb*), neutrophils (*N*), lymphocytes (*L*), monocytes (*M*), eosinophils (*E*), basophils (*B*), and platelets$\times 10^{-3}$ (*Plt*). If a patient were diagnosed as suffering from co morbidities, the appropriate concomitant therapies were immediately given and in principle ART was administered within three months of the morbidity diagnosis.

Therefore it is understood that examination records three or more months after concomitant therapies did not exist in the sample in the statistical analysis, both the CD4 and white blood cell counts were log-transformed. The researchers first obtained a regression equation for predicting CD4 counts from blood cell counts, by applying a stepwise linear regression to a sample of clinical examination records. This sample will be referred to as the "modeling sample."

Next, the researchers applied the regression equation to the records for four subsequent time periods following the first examination: 1 month (15-45 days), 3 months (60-120 days), 6 months (150-210 days), and 12 months (330-400 days). These four samples will be referred to as the "trial samples." They compared the predicted and actual CD4 counts of each record for each time period. The method of prediction using a regression equation obtained from the modeling sample has been widely applied in various fields.

This method will be referred to as the "sample-based prediction." The sample-based prediction method, however, did not account for individual variability. To this end, we developed a prediction method termed "individual-based prediction method". This method uses the same regression coefficients as the sample based regression equation, but instead of an average of the CD4 counts used in the sample-based prediction, it uses the previous CD4 count of each subject as the constant term. To summarize and compare the predictive performance of the prediction methods, we grouped CD4 counts into three categories; =200 cells/mm^3, 201-350 cells/mm^3 and >350 cells/mm^3."

The result and conclusion reached is as follows:

"**Results:** The sensitivity and the predictive-positive rate of CD4 counts below 350 cells/mm^3 were 89% and 88%, respectively, at a one-month interval and 76% and 84%, respectively, at a six-month interval."

"**Conclusion:** The individual-based prediction method of the CD4 count will be useful and critical in resource-poor regions where CD4 counts are not repeatedly available."

CHAPTER THREE

DATA PREPROCESSING AND MODEL SELECTION

3.1 Data preprocessing

Much of the raw data contained in databases is un-preprocessed, incomplete, and noisy. For example, the databases may contain: fields that are obsolete or redundant, missing values, outliers, data in a form not suitable for data mining models, values not consistent with policy or common sense.

To be useful for data mining purposes, the databases need to undergo preprocessing, in the form of data cleaning and data transformation. Data mining often deals with data that hasn't been looked at for years. So that much of the data may contains field values that have expired, are no longer relevant, or are simply missing. The overriding objective is to minimize GIGO: to minimize the "garbage" that gets into our model so that we can minimize the amount of garbage that our models give out. (Larose, 2005)

Scholars on data miming estimates that data preparation alone accounts for 60% of all the time and effort expanded in the entire data mining process. The preprocessing phase in this study took almost 75% of the time. This chapter describes data and examined principal methods for preparing the data to be mined, data cleaning, and handling missing values.

3.2 Data Description

The data sources for this research are Jimma, and Bonga hospitals ART data base. The format of Jimma hospital database is SQL server database management (DBMS). Bonga hospital uses Microsoft Access at the back end. Both databases have a front end to enter the data which is built by Visual Basic software.

The total data of the two databases is 8438 obtained from 2003 to 2012 and the numbers of cases are 6242 and 2197 in Jimma and Bonga hospitals respectively.

The name of file selected from Bonga database is called DataMart. DataMart file contains all the attributes necessary for the research. The Jimma database contains files called Patient and Register. The Patient and Register files in combination contain all the necessary attributes for the study. Those files are connected by patient card number to get a complete record. All the records in the database contain the name of the hospital, and patient card number to identify individual patient. Attributes from those file are the following: RegistrationDate, Sex, Age, ReligionID, MaritalStatusID, EducationalLevelID, Occupation, ARTStatus, Functional Status, ELDate, EligibleReasonID, ARTStartDate, FamilyPlanning, PregnantYN, OAWeight, OAWHO stage, CurrentRegimen, PastARVTreatment, and OACD4.

Table 3.1 Attribute description

No	**Attributes**	**Meaning**	**Value**	**Data type**
1	Sex	Sex of the patient	Female Male	Nominal
2	Age	Age of the patient	Numeric age values ranged as 0-14, 15-24, 25-49, 50-64	Numeric
3	Religion	The religion of the patient	1-Muslim 2-Orthodox 3-Protestant 4-Catholic 5-Other	Nominal
4	Marital Status ID	The marital status of the patient	1-Never married 2-Married(inc.de facto) 3-Separated 4-Divorsed 5-Widow/Widower	Nominal

5	Educational Level ID	Educational level of the patient	1-No education 2-Primary 3-Secondary 4-Teritiary	Nominal
6	ART Status	Status of ART care	OA-On taking ARV drug IN-In care for other disease EL-Eligible to be on ARV Drug ER-Eligible and ready to start ARV drug	Nominal
7	Functional Status	Functional level of the patient	W-working A-Ambulatory B-Bedridden	Nominal
8	Reason Eligible For ART	The reason for the patient eligible for ART	1-Clinical only 2-TLC 3-CD4 4- Transfer in(TI) 5-Clinical and TLC 6- Clinical and CD4	Nominal
9	ART Start Date	The date ART starts	From 2003-2012	Date
10	Family Planning	This attribute asks the usage of Family planning methods	Yes/No	Yes/No
11	Pregnant	Asks the patient pregnant or not on ART	Yes/No	Yes/No
12	OA Weight	The weight of the patient on ART	Numeric values ranged as:0-99,100-199, 200-349, 350-999,1000-2999	Numeric
13	OAWHO stage	WHO stages at which the patient is on ART	1-stage 1 2-sage 2 3-stage 3	Nominal

			4-stage 4	
14	Current Regimen	The current regimen the patient is taking on ART	1a(30)=d4T(30)-3TC-NVP 1a(40)=d4T(40)-3TC-NVP 1b(30)=d4T(30)-3TC-EFV 1b(40)=d4T(40)-3TC-EFV 1c=AZT-3TC-NVP 1d=AZT-3TC-EFV 2a, 2b, 2c, 2d. 4a, 4b, 4c and 4d. 5a, 5b, 5c, 5d Other	Nominal
15	Past ARV Treatment	Dose the patient took Past ARV Treatment	Yes for treatment No for not treatment	Yes/no
16	OACD4	The CD4 count of the patient currently on the ART	None zero positive number ranged as:0-49,50-99,100-199,200-349,350-999	Numeric

3.3 Statistical description of attributes

3.3.1 Sex

The sex attributes describes the sex of the patient as Female and Male. Out of the total 8438 data, 5471 are females and 2536 are males. The missing value of the data is 431 and it holds 5.1% of the total data.

Table3.2: Statistical summary of Sex attribute

Sex: Nominal data type		
Value	Frequency	Percent
Missing	431	5.1
Female	5471	64.8
Male	2536	30
Total	8438	100

3.3.2 Age

The age attribute contains the ages of the patient starts from 0 to above 65. This attribute has grouped into ranges of different values originally in the data base as follows 0-14, 15-24, 25-49, 50-64, and above 65. Out of the total data there are 32 missing values, which is 0.38% of the total.

Table3.3: Statistical summary of Age attribute

Age: Numeric data type		
Value	Frequency	Percent
Missing	32	0.38
0-14	290	3.44
15-24	2191	26
25-49	5692	67.45
50-64	231	2.73
Above 65	2	0.02
Total	8438	100

3.3.3 Religion

The religion attribute contains the religion of the patient is following. The data originally code as numeric data from 1 to 5. Based on the ministry of Health guide line the researcher recode the numeric codes to Muslim, Orthodox, Protestant, Catholic, Other in 1, 2,3,4,5 respectively. The missing values of the data are 535, which is 6.33% of the total data.

Table3.4: Statistical summary of Religion attribute

Religion: Nominal data type		
Value	Frequency	Percent
Missing	535	6.33
Muslim	2295	27.19
Orthodox	4489	53.19
Protestant	1012	11.99
Catholic	45	0.53
Other	59	0.69
Total	8438	100

3.3.4 Marital Status

The marital status attribute shows the marriage status of the patient. The data originally coded as numeric data from number 1 to number 5. Based on the Ministry of Health guide line the researcher recodes the numeric codes to nominal values of Never married, Married (inc.de facto), Separated, Divorced, Widow in 1, 2,3,4,5 respectively.

Table3.5: Statistical summary of Marital Status attribute

Marital Status : Nominal data type		
Value	Frequency	Percent
Missing	493	5.84
Never married	1681	19.91
Married(inc.de facto)	3793	44.94
Separated	477	5.65
Divorced	1010	11.96
Widow	984	11.66
Total	8438	100

3.3.5 Educational Level

The educational level attribute discuss on the level of education the patient has acquired. Originally the data coded as numeric data from 1 to 4. Based on the Ministry of Health guide line the researcher recodes the numeric code to nominal values of No education, Primary, Secondary, and Tertiary in 1,2,3,4 respectively. The data contains 501 missing values, which accounts 5.93% of the total data.

Table3.6: Statistical summary of Educational Level attribute

Educational Level : Nominal data type		
Value	Frequency	Percent
Missing	501	5.93
No education	1538	18.22
Primary	2998	35.52
Secondary	2786	33.01
Tertiary	613	7.26
Total	8438	100

3.3.6 ATR status

The attribute ART status shows the status of the patient condition currently. This attribute has five nominal data values, those are OA-On taking ARV drug, IN-In care for other disease, EL-Eligible to be on ARV Drug, ER-Eligible and ready to start ARV drug. The attribute do not have missing values.

Table3.7: Statistical summary of ART status attribute

ART status : Nominal data type		
Value	Frequency	Percent
Missing	0	0
OA-On taking ARV drug	6832	80.95
IN-In care for other disease	1586	18.79
EL-Eligible to be on ARV Drug	20	0.23
ER-Eligible and ready to start ARV drug	0	0
Total	8438	100

3.3.7 Functional status

The functional status attribute shows the patients status of physical and mental well being on the ART. The attribute contains three nominal data values coded as W for working, A for ambulatory, and B for bedridden. The data contains 1644 missing values, which is 19.48% of the total data.

Table3.8: Statistical summary of Functional status attribute

Functional status : Nominal data type		
Value	Frequency	Percent
Missing	1644	19.48
W-working	50900	60.31
A-Ambulatory	1357	16.08
B-Bedridden	347	4.11
Total	8438	100

3.3.8 Reason eligible for ART

Eligibility reason attribute asks why the patient is eligible to the treatment of ART. This attribute has nominal values to be returned as reasons those are: Clinical only, CD4, TLC, Transfer in (TI), Clinical and TLC, Clinical and CD4. The data originally entered as numeric values 0 to 6. Based on the Ministry of Health guide line the researcher recoded the numeric values 1 to Clinical only, 2 to TLC, 3 to CD4, 4 to Transfer in (TI), 5 to Clinical and TLC and 6 to Clinical and CD4.

Table3.9: Statistical summary of Reason eligible for ART attribute

Reason eligible for ART: Nominal data type		
Value	Frequency	Percent
Missing	2751	32.6
Clinical only	927	11.00
TLC	386	4.57
CD4	3683	43.6
Transfer in(TI)	136	1.6
Clinical and TLC	341	4.04
Clinical and CD4	214	2.53
Total	8438	100

3.3.9 ART Start Date (year)

The start year derived from the ART start date of the patient. The year attribute describes the year the patient starts ART and it starts from the year 2003 to 2012. The data contains 2127 missing values, which accounts for 25.2% of the total data.

Table3.10: Statistical summary of ART start Date (year) attribute

ART start Date (year): Numeric data type		
Value	Frequency	Percent
Missing	2127	25.20
2003	41	0.49
2004	158	1.83
2005	1091	12.92
2006	2678	31.73
2007	1344	15.92
2008	297	3.50
2009	341	4.02
2010	170	2.01
2011	190	2.25
2012	1	0.01
Total	8438	100

3.3.10 Family Planning

This attribute indicates that the usage of family planning method of the patient. Originally the data coded as 0 for not using and 1 for using methods of family planning. The researcher recoded the 0's and 1's code to No and Yes codes. The data contains 1196 missing values of the total data, which is 14.17% of the total data.

Table3.11: Statistical summary of Family planning attribute

Family planning: Nominal data type		
Value	Frequency	Percent
Missing	1196	14.17
Yes	2590	30.69
No	4652	55.12
Total	8439	100

3.3.11 Pregnant

The attribute pregnant indicates that whether the patient is pregnant or not. Originally the data coded as Yes and No for being pregnant and not pregnant respectively. The data do not have missing value. In this data there is no patient who is pregnant.

Table3.12: Statistical summary of Pregnant attribute

Pregnant: Nominal data type		
Value	Frequency	Percent
Missing	0	0
Yes	0	0
No	8438	100
Total	8438	100

3.3.12 OA Weight

The OA weight attribute shows the weight of the patient on the ART service. The weight of the patient in the data starts from 3k.g to above 101k.g. This attribute has not grouped into ranges of different values originally in the database. In order to group the data in to different intervals, the researcher decides first to replace the missing values by mean weight and then group into different intervals by using discritization method. Section 3.4 discusses the discritization of the OA weight attribute. Table 3.13 shows the statistically summary of the weight attribute.

Table3.13: Statistical summary of OA weight attribute

OA weight: Numeric data type		
Value	Frequency	Percent
Missing	2541	31.11
Mean	48.5	
Mode	50	
Minimum	3	
Maximum	101	

3.3.13 WHO stage

The WHO stage attribute contains the stages of the patient in the treatment. The data originally coded as I, II, III, IV the researcher recode this attribute value in Stage 1, Stage 2, Stage 3 and Stage 4 respectively. The data contains 2313 missing values, which is 27.4% of the total data.

Table3.14: Statistical summary of WHO stage attribute

WHO stage: Nominal data type		
Value	Frequency	Percent
Missing	2313	27.40
Stage 1	393	4.65
Stage 2	1442	17.08
Stage 3	2971	35.20
Stage 4	1319	15.63
Total	8438	100

3.3.14Current regimen

Current regimen attribute describes the regimen type of the patient is taking. The regimen contains two line regimens firs line and second line. The lines also classifies as first line and second line adult and firs line and second line child. First line adult includes 1a (30) or d4T (30)-3TC-NVP, 1a (40) or d4T (40)-3TC-NVP, 1b (30) or d4T (30)-3TC-EFV, 1b (40) or d4T (40)-3TC-EFV, 1c or AZT-3TC-NVP and 1d or AZT-3TC-EFV. Second line adult includes 2a, 2b, 2c, 2d. The first line child includes 4a, 4b, 4c and 4d. Second line child include 5a, 5b, 5c, 5d and other. The data contains 2147 missing values, which is 25.44% of the total data.

Table3.15: Statistical summary of Current regimen attribute

Current regimen: Nominal data type		
Value	Frequency	Percent
Missing	2147	25.44
1a(30)=d4T(30)-3TC-NVP	3469	41.11
1a(40)=d4T(40)-3TC-NVP	755	8.94
1b(30)=d4T(30)-3TC-EFV	406	4.81
1b(40)=d4T(40)-3TC-EFV	124	1.46
1c=AZT-3TC-NVP	781	9.25
1d=AZT-3TC-EFV	192	2.27
2a	-	
2b	-	
2c	-	
2d	-	
4a	37	0.43
4b	9	0.11

4c	172	2.03
4d	17	0.20
5a	-	
5b	-	
5c	-	
5d	-	
Other	329	3.89
Total	8438	100

3.3.15 Past ARV treatment

This attribute indicates that whether the patient had took or not any ARV treatment in the past. Originally the data coded as Yes and No for taken treatment and No for not taken treatment in respectively. The data contains 311 missing values, which accounts 3.68% of the total data.

Table3.16: Statistical summary of Past ARV treatment attribute

Past ARV treatment: Nominal data type		
Value	Frequency	Percent
Missing	311	3.68
Yes	780	9.23
No	7347	87.03
Total	8438	100

3.3.16 OA CD4count

OA CD4 count describes the CD4 count values of the patient on ART. The CD4 count already grouped in the database as 0-49, 50-99, 100-199, 200-349, 350-999 and 1000-2999. The data contains 661 missing values which are 7.83% of the total data. Replacing the missing value is discussed in section 3.4.1. The numeric value of the CD4 count has transformed in to nominal values based on the base line guide of the treatment. The transformation of the value is discussed in section 3.4.2.

Table3.17: Statistical summary of OA CD4 count attribute.

OA CD4 count: Numeric data type		
Value	Frequency	Percent
Missing	661	7.83
0-49	3280	38.87
50-99	1060	12.56
100-199	1987	23.54

200-349	1286	15.24
350-999	151	1.78
1000-2999	14	0.16
Total	8438	100

3.4 Data cleaning

Many of the existing data are not clean; they need the effort of the researcher in the area to get rid of bad data. Most Data cleaning tasks include in handling missing values, smoothing noisy data, identifying and removing outliers, and resolving inconsistencies.

3.4.1 Handling Missing Values

In many real world applications of data mining, the subset of cases with complete data may be relatively rare. Available samples and also future cases may have values missing. Some of the data mining methods accept missing values and satisfactorily process data to reach a final conclusion. Other methods require that all values be available. The question is whether these missing values can be filled in during data preparation, prior to the application of the data mining methods. The simplest solution for this problem is the reduction of the data set and the elimination of all samples with missing values. That is possible when large data sets are available, and missing values occur only in a small percentage of samples. If the researcher does not drop the samples with missing values, then he/she have to find values for them (Kantardzic, 2003). In this study the researcher intended not to cut the data set with missing values, rather find possible solutions.

There are known practical solutions for handling missing values. One of the solutions is a discussion with the domain expert. The discussion of the data miner with the expert can manually examine samples that have no values and enter a reasonable, probable, or expected value, based on a domain experience. The method is straightforward for small numbers of missing values and relatively small data sets. But, if there is no obvious or plausible value for each case, the miner is introducing noise into the data set by manually generating a value (Kantardzic, 2003). Since discussion with the domain experts works good for small number of missing values and the missing values in this study are different from attributes to attributes the researcher find another solution.

The approach that gives an even simpler solution for elimination of missing values in this study is a formal, often automatic replacement of missing values with some constants. su Replace all missing values with a single global constant (a selection of a global constant is highly application-dependent). The first type of replacement is replacing a missing value with its mean for numeric data type. The other replacement is done by replacing a missing value with mode for the nominal data type. For example the sex attribute has 5.1% missing value and the modal value for the attribute is female then, the researcher decides to replace with female value for the missed attributes. Another example OA weight attribute's missing value replaced with the following procedure: calculate the mean value of the weight and the result value is 48.5, this value replace the missed value.

Table 3.18 Summary of missing value handled

No	Attributes	Percentage of missing values	Replaced with	Data type	Remark
1	Sex	5.1	Female	Nominal	Mode
2	Age	0.38	28.8(25-49)	Numeric	Mean
3	Religion	6.33	Orthodox	Nominal	Mode
4	Marital status	5.84	Married	Nominal	Mode
5	Educational level	5.93	Primary	Nominal	Mode
6	Functional Status	19.48	W-working	Nominal	Mode
7	Reason Eligible For ART	1.25	Transfer in (TI)	Nominal	Mode
8	ART start Date	25.20	2006	Numeric	Mode
9	Family Planning	14.14	No	Yes/no	Mode
10	OA weight	30.11	48.5	Numeric	Mean
11	WHO stage	27.4	Stage3	Nominal	Mode
12	Current Regimen	34.83	1a(30)	Nominal	Mode
13	Past ARV treatment	3.68	No	Yes/no	Mode
14	OACD4	7.83	102	Numeric	Mean

3.4.2 Data transformation

Data transformation converts data from different sources into common new format. Apply data reduction & data categorization/binning to ease data mining. Discretization: Reduce data size by dividing the range of a continuous attribute into intervals. Interval

labels can then be used to replace actual data values. The Discretization is performed by method of Equal-width (distance) partitioning. This method divides the range into N intervals of equal size. The method assign A and B as the lowest and highest values of the attribute. The width of intervals: W = (B –A)/N. Applying the formula the OA weight attribute discretize as shown in table 3.19.

Table 3.19 Discretized result of weight attributes.

Label	Frequency	Percent
3-27	280	3.31
27.5-52	3415	40.47
52.5-77	2170	25.71
77.5-102	31	0.36
Missing	2541	30.11
Total	8438	100

The other transformation done in this research is transforming the numeric value of OA CD4 count to nominal value. In order to transform the value the researcher follows the baseline treatment guide. Different scholars suggest that there should be a cut point for CD4 count to be low and normal. Mellors, (1997) classifies as: if CD4 count is below 200 cells/mm^3 it can be categorized to low CD4 status. Where as, if the CD4 count is above 200 cells/mm^3 it can be categorized to normal CD4 status. Table 3.20 below shows the transformed OA CD4 count.

Table 3.20 Transformed result of OA CD4 count

OA CD4 count: Nominal data type	
Value	Frequency
Low	6987
Normal	1451
Total	8438

3.5 Model implementation

In this study for building classification model decision tree and PART are used. The next section briefly discusses about the decision tree algorithm C4.5 and J48.

3.5.1 Decision trees

Decision trees are a way of representing a series of rules that lead to a class or value. A particularly efficient method for producing classifiers from data is to generate a decision tree. The decision-tree representation is the most widely used logic method by navigating the decision tree. It is possible to classify a case by deciding which branch to take, starting at the root node and moving to each subsequent node until an output class is reached. Each node uses the data from the case to choose the appropriate branch. Decision trees are grown through an iterative splitting of data into discrete groups, where the goal is to maximize the "distance" between groups at each split. (Two crows corporation, 1999)

A typical decision-tree learning system adopts a top-down strategy that searches for a solution in a part of the search space. A decision tree consists of nodes that where attributes are tested. The outgoing branches of a node correspond to all the possible outcomes of the test at the node. A simple decision tree example is adopted from (Kantardzic, 2003).

Samples with two input attributes X and Y are given in Figure 3.1. All samples with feature values X>1 and Y=B belong to Class2, while the samples with values X<1 belong to Class1, whatever the value for feature Y. The samples, at a non leaf node in the tree structure, are thus partitioned along the branches and each child node gets its corresponding subset of samples.

Decision trees which are used to predict categorical variables are called classification trees because they place instances in categories or classes. Decision trees used to predict continuous variables are called regression trees. Decision trees make few passes through the data (no more than one pass for each level of the tree) and they work well with many independent variables. As a consequence, models can be built very quickly, making them suitable for large data sets. Decision trees that use univariate splits have a simple representational form, making it relatively easy for the user to understand the inferred model; at the same time, they represent a restriction on the expressiveness of the model. In general, any restriction on a particular tree representation can significantly restrict the functional form and thus the approximation power of the model. (Kantardzic.M.2003)

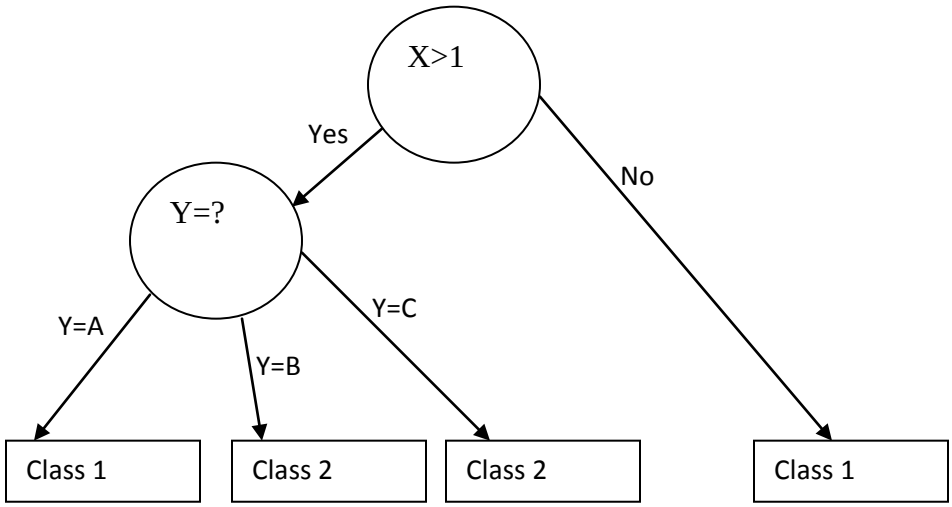

Fig 3.1 Simple decision tree source: Kantardzic, 2003

Decision trees handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and explosion of independent variables inherent in neural nets. Some classification trees were designed for and therefore work best when the independent variables are also categorical. Continuous predictors can frequently be used even in these cases by converting the continuous variable to a set of ranges (binning). Some decision trees do not support continuous output variables (i.e., will not build regression trees), in which case the dependent variables in the training set must also be binned to output classes (Two crows corporation, 1999)

3.5.1.1 C4.5 Algorithm: Generating a decision tree

The most important part of the C4.5 algorithm is the process of generating an initial decision tree from the set of training samples. As a result, the algorithm generates a classifier in the form of a decision tree; a structure with two types of nodes: a leaf, indicating a class, or a decision node that specifies some test to be carried out on a single-attribute value, with one branch and sub tree for each possible outcome of the test.

A decision tree can be used to classify a new sample by starting at the root of the tree and moving through it until a leaf is encountered. At each non leaf decision node, the features' outcome for the test at the node is determined and attention shifts to the root of the selected subtree. For example, if the classification model of the problem is given with the

decision tree in Figure 3.2 then the algorithm will create the path through the nodes A, C, and F (leaf node) until it makes the final classification decision: *CLASS2*.

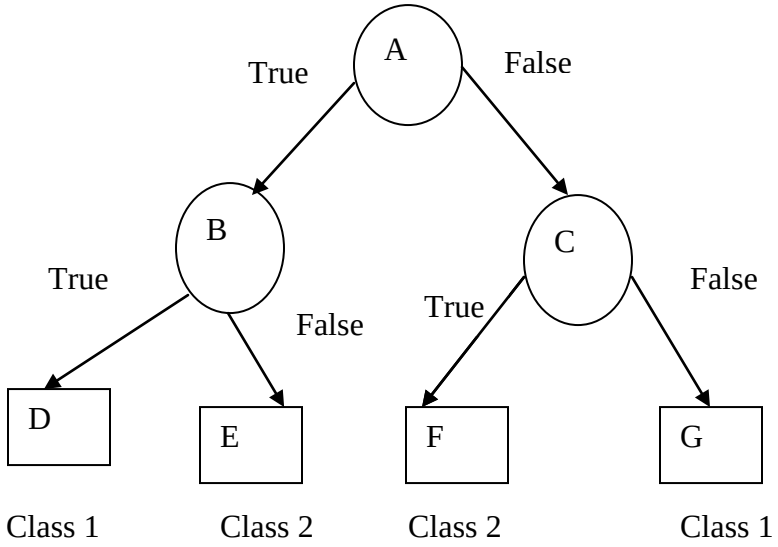

Fig 3.2 classification of new sample based on decision tree model source: Kantardzic, 2003

The skeleton of the C4.5 algorithm is based on Hunt's *CLS* method for constructing a decision tree from a set T of training samples. Let the classes be denoted as $\{C_1, C_2\ldots, C_k\}$. There are three possibilities for the content of the set T:

- T contains one or more samples, all belonging to a single class C_j. The decision tree for T is a leaf identifying class C_j.
- T contains no samples. The decision tree is again a leaf but the class to be associated with the leaf must be determined from information other than T, such as the overall majority class in T. The C4.5 algorithm uses as a criterion the most frequent class at the parent of the given node.
- T contains samples that belong to a mixture of classes. In this situation, the idea is to refine T into subsets of samples that are heading towards a single-class collection of samples. Based on single attribute, an appropriate test that has one or more mutually exclusive outcomes {O1, O2…,On} is chosen. T is partitioned into subsets T1, T2…,Tn where Ti contains all the samples in T that have outcome Oi

of the chosen test. The decision tree for T consists of a decision node identifying the test and one branch for each possible outcome (examples of this type of nodes are nodes A, B, and C in the decision tree in Figure 3.2a).

The same tree-building procedure is applied recursively to each subset of training samples, so that the i^{th} branch leads to the decision tree constructed from the subset Ti of training samples. The successive division of the set of training samples proceeds until all the subsets consists of samples belonging to a single class.

The tree-building process is not uniquely defined. For different tests, even for a different order of their application, different trees will be generated. Ideally, we would like to choose a test at each stage of sample-set splitting so that the final tree is small. Since we are looking for a compact decision tree that is consistent with the training set, why not explore all possible trees and select the simplest?

Unfortunately, the problem of finding the smallest decision tree consistent with a training data set is NP-complete. Enumeration and analysis of all possible trees will cause a combinatorial explosion for any real-world problem. For example, for a small database with five attributes and only twenty training examples, the possible number of decision trees is greater than 106, depending on the number of different values for every attribute. Therefore, most decision tree construction methods are non-backtracking, greedy algorithms. Once a test has been selected using some heuristics to maximize the measure of progress and the current set of training cases has been partitioned, the consequences of alternative choices are not explored. The measure of progress is a local measure, and the gain criterion for a test selection is based on the information available for a given step of data splitting.(Kantardzic, 2003)

3.5.1.2 Attribute selection measure

The original ID3 algorithm used a criterion called gain to select the attribute to be tested which is based on the information theory concept: entropy. The following relation gives the computation of the entropy of the set S (bits are units):

Suppose we have the task of selecting a possible test with n outcomes (n values for a given feature) that partitions the set T of training samples into subsets T1, T2, …, Tn.

The only information available for guidance is the distribution of classes in T and its subsets Ti. If S is any set of samples, let freq (Ci, S) stand for the number of samples in S that belong to class Ci (out of k possible classes), and let |S| denote the number of samples in the set S.

$$Info(S) = -\sum_{i=1}^{k}((\text{freq}(\text{Ci}, \text{S})/|\text{S}| * log2(freq(Ci, S)/|S|))$$

Now consider a similar measurement after T has been partitioned in accordance with n outcomes of one attribute test X. The expected information requirement can be found as the weighted sum of entropies over the subsets:

$$Infox(T) = -\sum_{i=1}^{n}((|\text{Ti}|/|\text{T}|) * Info(Ti)$$

The quantity

$$\text{Gain(X)} = \text{Info (T)} - \text{Infox (T)}$$

Measures the information that is gained by partitioning T in accordance with the test X. The gain criterion selects a test X to maximize Gain (X), i.e., this criterion will select an attribute with the highest info-gain.

3.5.1.3 C4.5 Algorithm: Generating decision rules

Large decision trees are difficult to understand because each node has a specific context established by the outcomes of tests at antecedent nodes. To make a decision-tree model more readable, a path to each leaf can be transformed into an IF-THEN production rule. While the IF part consists of all tests on a path, and the THEN part is a final classification. Rules in this form are called decision rules, and a collection of decision rules for all leaf nodes would classify samples exactly as the tree does. As a consequence of their tree origin, the IF parts of the rules would be mutually exclusive and exhaustive, so the order of the rules would not matter. The following example is adopted from (Kantardzic.M.2003) as follows:

An example of the transformation of a decision tree into a set of decision rules is given in Figure 3.3(Kantardzic, 2003), where the two given attributes, A and B, may have two possible values, 1 and 2, and the final classification is into one of two classes.

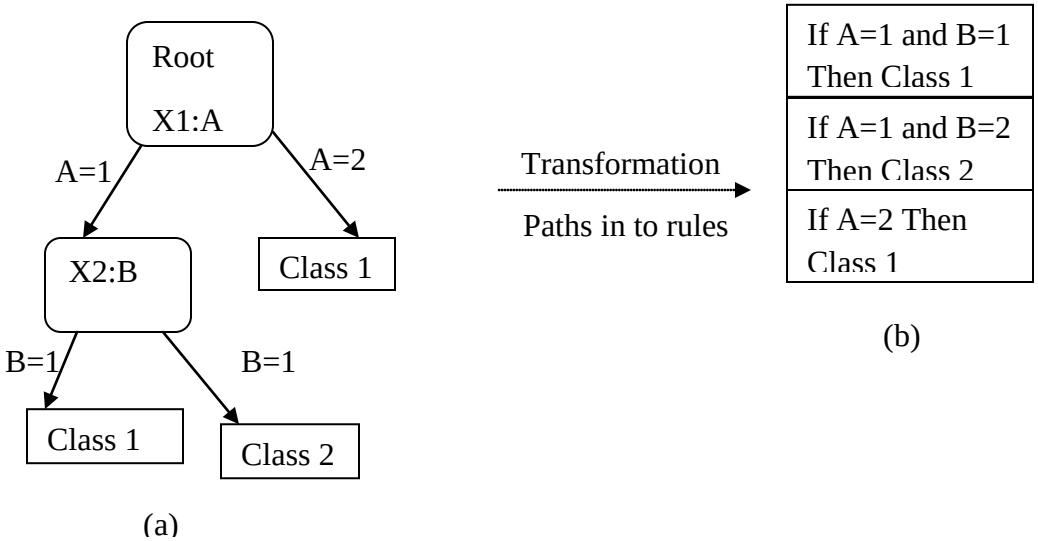

Fig 3.3 classification of new sample based on decision tree model: (a) Decision Tree; (b) Decision rules

3.5.2 J48 Decision tree

The algorithm used by Weka is known as J48. J48 is a version of an earlier algorithm developed by J. Ross Quinlan, the very popular C4.5. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data the researcher also decides to use J48 in this study.

It is important to understand the variety of options available when using this algorithm, as they can make a significant difference in the quality of results. In many cases, the default settings will prove adequate, but in others, each choice may require some consideration.

The J48 algorithm gives several options related to tree pruning. Many algorithms attempt to "prune", or simplify, their results. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential over-fitting. The basic algorithm described above recursively classifies until each leaf is pure,

meaning that the data has been categorized as close to perfectly as possible. This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular idiosyncrasies of that data. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data. This is because pruning employs various means to relax the specificity of the decision tree, hopefully improving its performance on test data. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy (Wittenet.al, 2005).

J48 employs two pruning methods. The first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub tree rising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub tree rising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub tree raising can be somewhat computationally complex (Han et.al, 2006).

CHAPTER FOUR

EXPERIMENT AND ANALYSIS OF CLASSIFICATION MODEL

4.1 Attribute ranking

As it has been discussed in chapter 3 on section 3.5.1.2 about Information Gain, attribute selection is very important in building decision tree model. By considering the importance of selecting the attribute, the researcher performed weka to select the best attribute. The implementation by Weka attribute ranking filter using information gain uses attribute Evaluator in supervised class of nominal values of 16 attributes.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 16 OACD4):
        Information Gain Ranking Filter

Ranked attributes:
 0.1195714   8 EligibleReason
 0.0542675   6 ARTStatus
 0.0352196   9 ARTStartYear
 0.033732   12 OAWeight
 0.0199281  13 OAWHOstage
 0.0099687  14 CurrentRegimen
 0.0086814  10 FamilyPlanning
 0.0054349   7 FunctionalStatus
 0.0023818   4 MaritalStatus
 0.0015243  15 PastARV
 0.0005472   3 Religion
 0.0001696   2 Age
 0.0001688  11 Pregnant
 0.0001665   1 Sex
 0.0000629   5 EducationalLevel

Selected attributes: 8,6,9,12,13,14,10,7,4,15,3,2,11,1,5 : 15
```

Fig 4.1: Result of ranking attribute

As it can be seen from the result output fig 4.1 of attribute selection using entropy based information gain method of Weka, the top 10 determining attributes of the data set for predicting CD4 status of the patient on ATR care service are: EligibleReasaon, ARTStatus, ARTStartyear, OAWeight, OAWHOstage, CurrentRegimn, FamiliyPlaning, FunctionalStatus, MaritalStatus, PastARV With gain of 0.120, 0.054, 0.035, 0.034, 0.020, 0.010, 0.009, 0.005, 0.002, 0.002 in respectively. The digit of the gain is rounded off to the nearest number.

The result included different expressive meanings such as the attribute information gain value, the attribute's indices in the dataset and the attribute name. The ranked attribute includes the attribute's indices number from top to least and the total attribute numbers used in the comparison. The rank of the attribute shows the relevance of the attributes to the experimentation by excluding the least relevant attributes.

4.2 Classification model building

In classification model building the researcher intended to build four scenarios with all attributes those are: Binary decision tree with pruning, Binary decision tree without pruning, generalized decision tree with pruning, generalized decision tree without pruning with all attributes for all scenarios.

The researcher also intended to build a model based on the reduced (10 top attributes). Building a model after reduction of attributes doubles the number of scenarios into eight scenarios. The scenarios for reduced attributes are the same to the above experiment. The difference between the two experiments is, in reduced attribute experiment the number of attributes decreased to 10 based on the information gain as it has been discussed in section 4.1.

The purpose of building a tree model is to get the smallest tree that has purest leaf nodes. To say a node is purer leaf node, if the model precise its classification is. It is known that a leaf node with all the instances in it are correctly classified is clearly better than from the one with half and below half correctly classified instances. The better and worseness of the different algorithm varies from dataset to dataset (Han et.al, 2006).

Since algorithm selection is important to build the classification model, the researcher tried to implement decision tree model with prune. The pruning can be take place by splitting the data into sub samples that is more pure than the original sample data. As Kamber and Han (2006) discussed, the ideal situation is when each sub-sample consists of instances that have the same value for the class attribute i.e. completely pure nodes.

For a matter of this study the word purity is referred to how similar (homogenous) the sub samples are in relation to the class variables. To build the classification tree model the researcher intended to decide on the portion of the dataset, to be used for training and testing purpose. In order to get the amount of sample used for training and testing the researcher computes different experiments by changing the percentage of the dataset. The experiment of this portion can be seen in section 4.2.2. The figure below shows the Weka Explorer to use decision tree classification using J48.

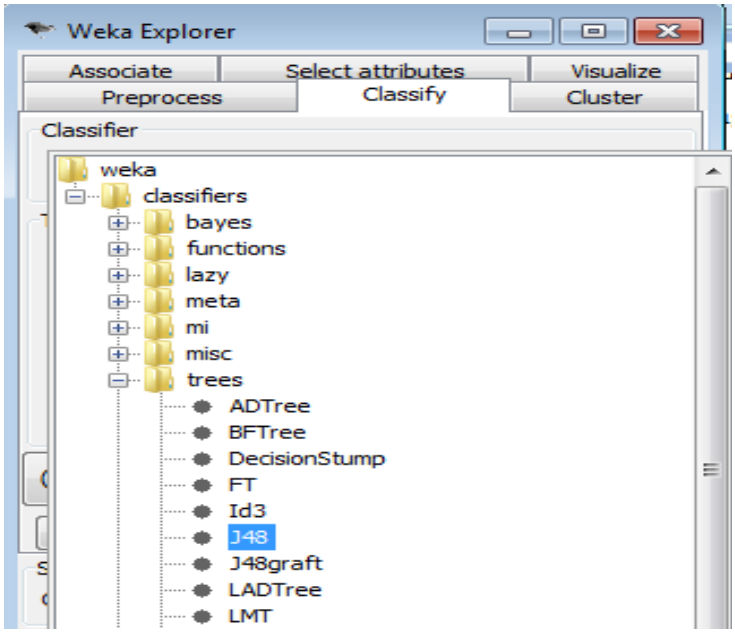

Fig 4.2 Classification of tree using J48

4.2.1 Confidence factor and incorrectly classified instances

Confidence factor has impact on the percentage of classifying instances in to correctly and incorrectly. The researcher conducted different experiments by changing the confidence factor values incrementally. As it can be seen from the experiment result table when the confidence factor increases, the incorrectly classified instance decreases. The researcher decide to take confidence factor 0.5 as a confidence factor for all the experiments, this is because confidence factor 0.5 has minimum incorrectly classified instances than the other experiments. The Table 4.1 below show that the pattern of confidence factors and incorrectly classified instances.

Table 4.1: Incorrectly classified instance for different confidence factors

Confidence factor	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.4	0.45	0.5	0.55
Incorrectly classified instances	988	985	976	967	903	909	879	867	854	823	824	800

4.2.2 Measuring Classifier accuracy for decision tree

Classifier accuracy is the important issue in classification model building. In measuring classifier accuracy on unseen data, there are different validation methods for decision tree. Types of validation methods are: classifying the data into training and test set i.e. full training set, K-fold cross validation (10-fold cross validation) in Weka.

According to Bramer, (2007) it is important to check the appropriateness of the dataset for selecting certain validation method. For the matter of selecting the validation of methods, it is necessary to see the appropriateness of the available dataset.

In order to get the minimum percent to be used as learning purpose the researcher computes different experiments by varying the percent of k-fold cross validation. For the purpose of this research the experiment starts by taking samples from 10% to 50% as it has been shown on the table 4.2. The precision of all of the experiment is nearly the same. Since there is no much difference on the precision of the experiments, the researcher decides to set the percentage of testing set into 50%.

Table 4.2 Precisions of k- fold cross validation

Sample %	Accuracy
10	0.89
20	0.90
30	0.89
40	0.9
50	0.9

4.2.2 Binary Decision Tree Model Building

One of the classification decision tree models has been built by the researcher is binary decision tree model. The binary decision tree split internal node branches exactly into two sub trees i.e. it will have two branches at each node (Bramer , 2007).

The model built by Binary decision tree have four scenarios those are: Binary decision tree with all attributes without pruning, Binary decision tree with all attributes with pruning, Binary decision tree with some selected attributes without pruning, Binary decision tree with some selected attributes with pruning. The figure bellows shows using J48 when binary classification is “True” this means the classifier is now using binary tree classification.

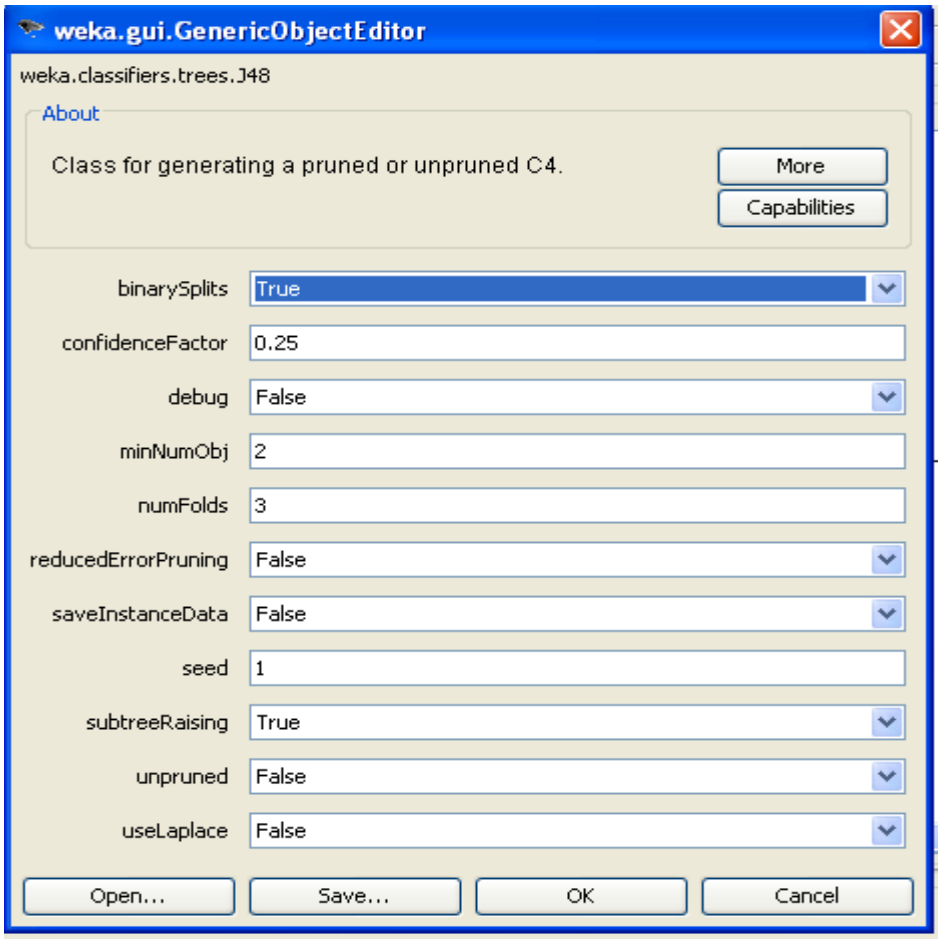

Fig 4.3 Binary decision tree classification

4.2.3 Generalized Decision Tree model Building

As it has been discussed in the section 4.2 generalized decision tree model is one of the experiments selected by the researcher to be implemented.

Generalized decision tree model is built by splitting into more than two sub trees. Even though it split more than two branches, they may be less important and may not have good knowledge. The model is built by setting the 'binarySplit' to 'false'.

The experiment built help for analyzing the comparison of the model with the other model (Binary). To build generalized decision tree model the setting of the Weka is changed to Binary split to "False" and the other setting of the Weka is remain the same to binary decision tree.

The model have four scenarios those are: Generalized decision tree with all attributes without prune, generalized decision tree with all attributes with prune, generalized decision tree with some selected attributes with prune, generalized decision tree with some selected attributes without prune. The prune and none prune done by setting the unpruned ’True’ for non prune one and unpruned to ‘False’ for prune one.

4.3 Experiment description

The experiment and analysis of classification models chapter concerns on experimenting different scenarios of classification models using J48 decision tree algorithm. The scenarios are done by using the selected confidence factor as stated in section 4.2.1 and setting the parameter into different values. The analysis of the model is concerned on selecting the best model from the experiments done. The selection of the model from the scenarios is based on comparing the result values of the analysis. The comparison parameters are number of leaves, tree size, time, Correctly classified instances(CCI), incorrectly classified instances(ICI), True positive rate(TP), False Positive rate(FP), Precision, Recall and ROC Area. The comparison of the result ended with selecting the best model of the eight scenarios done. The next experiment is conducting PART rule generating algorithm with the same value of the previous model of J48 selected. The result values of the PART rule and the selected J48 model compared for obtaining the best rule of experiment.

The experiment of the scenario includes: Run information which contains scheme, Relation, Instances, Attributes and Test mode. Classifier model contains number of leaves, size of the tree, time taken to build model. Summary with Correctly Classified Instances, Incorrectly Classified Instances and Total number of instances Detailed Accuracy by Class includes TP Rate, FP Rate, Precision, Recall, F-Measure ROC Area, Class and Confusion Matrix. The confusion matrix of the class which is a base for calculating accuracy measures of all scenarios.

4.4 J48 Algorithms model building

As it has discussed in the chapter four there are eight scenarios to be experimented for decision tree classification. The experiment is classified into two types based on the number of attributes. The first type of experiment includes four scenarios with all 16 attributes. Out of these four scenarios two of them are Binary decision tree with pruning and without pruning and the rest two are General decision tree with and without pruning.

The second type of experiment also has four scenarios those are similar with the first except this used reduced attributes based on the attribute ranking discussed at section 4.1.

The eight experiments for J48 decision tree scenarios are listed below:

Scenario #1Binary decision tree without pruning with all attributes
Scenario #2General decision tree without pruning with all attributes
Scenario #3Binary decision tree without pruning with reduced attributes
Scenario #4General decision tree without pruning with reduced attributes
Scenario #5Binary decision tree with pruning with all attributes
Scenario #6General decision tree with pruning with all attributes
Scenario #7Binary decision tree with pruning with reduced attributes
Scenario #8General decision tree with pruning with reduced attributes

Scenario #1Binary decision tree without pruning with all attributes

=== Run information ===

Scheme: weka.classifiers.trees.J48 -U -B -M 2

Relation: ART1

Instances: 8438

Attributes: 16

Sex

Age

Religion

MaritalStatus

EducationalLevel

ARTStatus

FunctionalStatus

EligibleReason

ARTStartYear

FamilyPlanning

Pregnant

OAWeight

OAWHOstage

CurrentRegimen

PastARV

OACD4

Test mode: split 50.0% train, remainder test

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

J48 unpruned tree

Number of Leaves : 649

Size of the tree : 1297

Time taken to build model: 0.42 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	3701	87.7222 %
Incorrectly Classified Instances	518	12.2778 %
Kappa statistic	0.5493	
Mean absolute error	0.1316	
Root mean squared error	0.3227	
Relative absolute error	46.2058 %	
Root relative squared error	85.9994 %	
Total Number of Instances	4219	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.934	0.403	0.919	0.934	0.927	0.836	low
	0.597	0.066	0.65	0.597	0.622	0.836	normal
Weighted Avg.	0.877	0.346	0.874	0.877	0.875	0.836	

=== Confusion Matrix ===

a	b	<-- classified as
3274	230	a = low
288	427	b = normal

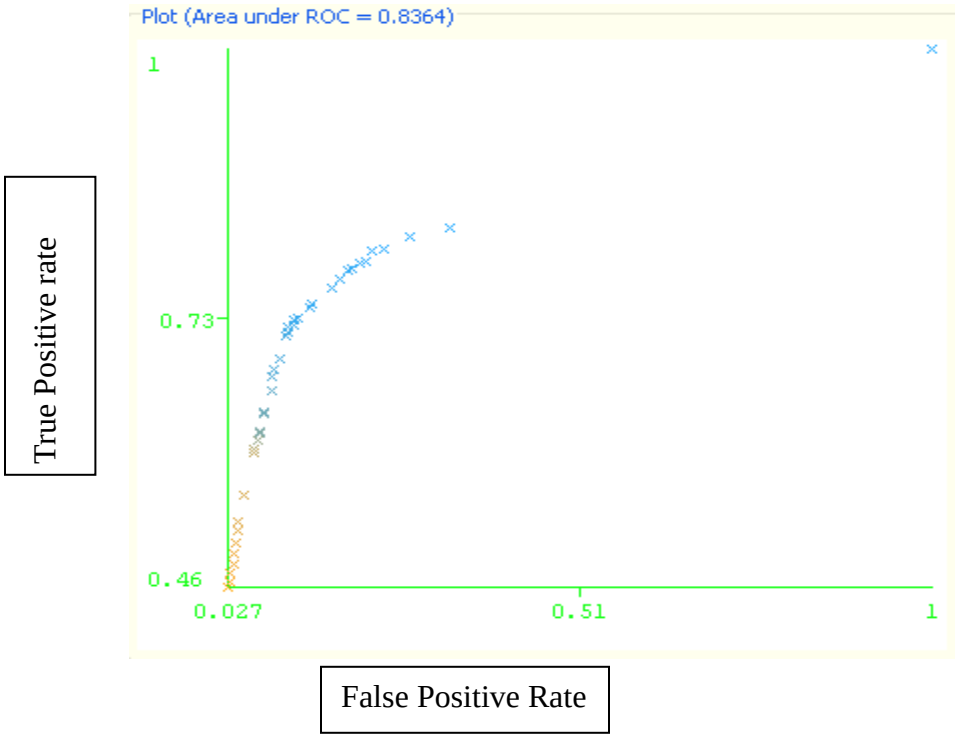

Fig 4.4 ROC Area curve for scenario #1

Table 4.3 Summary of Experiment #1

Type	Pruned	Leaves	Tree size	Time	CCI in %	ICI in %	TP Rate	FP Rate	Precision	Recall	ROC Area
Binary	No	649	1297	0.42	87.72	12.28	0.88	0.35	0.87	0.88	0.84

Result analysis of scenario #1

It is possible to see on the summary table of scenario #1, in table 5.1. The scenario has generated complex tree structure which is not possible to see. The time required to the experiment is 0.42 second. Classifying records correctly into both 'true' and 'false' is high (87.72) where as incorrectly classifying is low (12.28). True positive rate of the experiment is high (0.88) and false positive rate is low (0.35). Precision of the scenario is high (0.87). Recall of the experiment is high (0.88). The ROC Area is above 0.5, which is the minimum possible acceptable value of ROC curve. As it can be seen from the figure of ROC curve 0.84 is above the diagonal. Since the ROC area value is above the given minimum criteria, it is possible to say the model is highly accurate.

Scenario #2General decision tree without pruning with all attributes

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.932	0.369	0.925	0.932	0.928	0.859	low
	0.631	0.068	0.653	0.631	0.642	0.859	normal
Weighted Avg.	0.881	0.318	0.879	0.881	0.88	0.859	

=== Confusion Matrix ===

a	b	<-- classified as
3264	240	a = low
264	451	b = normal

Result analysis of scenario #2

The result of the experiment of the scenario is shown in table 5.2. As it can be seen from the table the tree size and leaves are more complex than experiment 1. The time required for the experiment is very shorter than experiment 1 it is (0.19). Also the CCI of this experiment is higher than experiment1. TP rate is the same as the previous. FP rate is low and lower than the previous. Precision is high and higher than experiment 1 which is

(0.88). The recall is the same (0.88). ROC area is high and enough to say accurate that is (0.86).

Scenario #3Binary decision tree without pruning with reduced attributes

=== Run information ===

Instances: 8438

Attributes: 11

MaritalStatus

ARTStatus

FunctionalStatus

EligibleReason

ARTStartYear

FamilyPlanning

OAWeight

OAWHOstage

CurrentRegimen

PastARV

OACD4

Test mode: split 50.0% train, remainder test

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.952	0.484	0.906	0.952	0.928	0.849	low
	0.516	0.048	0.687	0.516	0.589	0.849	normal
Weighted Avg.	0.878	0.41	0.869	0.878	0.871	0.849	

=== Confusion Matrix ===

a b <-- classified as

3336 168 | a = low

346 369 | b = normal

Result analysis of scenario #3

This scenario shows a less complicated tree than the above two scenarios which is 423 and 845 leaves and tree sizes respectively. The time is 0.19 second which is less time and similar to the previous experiment. Correctly Classified Instances are 87.82% which is less than the previous one. ICI is12.18 % which is higher than the previous experiment.TP rate is the same to the above experiments. FP rate is higher than that of two experiments. The precision is high (0.87). The recall is the same as the above experiments. ROC Area is high (0.85) and enough to say accurate.

Scenario #4General decision tree without pruning with reduced attribute

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.959	0.485	0.906	0.959	0.932	0.862	low
	0.515	0.041	0.722	0.515	0.601	0.862	normal
Weighted Avg.	0.884	0.41	0.875	0.884	0.876	0.862	

=== Confusion Matrix ===

a b <-- classified as

3362 142 | a = low

347 368 | b = normal

Result analysis of scenario #4

The decision tree of the experiment is complicated and not visible it has 1264 leaves and 1513 tree size. The time required for the experiment is less than the previous one. CCI of the experiment is 88.40 which is a little higher than the previous experiment. ICI of the experiment is 11.59 it is less than experiment 3.TP Rate is the same to the above experiments. FP rate is similar with the previous experiment. Precision of the experiment

is high (0.88). The recall of the experiment is the same to the above experiments.ROC area is high (0.86) which is enough to say accurate.

Scenario #5Binary decision tree with pruning with all attributes

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.946	0.438	0.914	0.946	0.93	0.797	low
	0.562	0.054	0.68	0.562	0.616	0.797	normal
Weighted Avg.	0.881	0.373	0.874	0.881	0.876	0.797	

=== Confusion Matrix ===

a	b	<-- classified as
3315	189	a = low
313	402	b = normal

Result analysis of scenario #5

In this experiment the leaves and tree size are less than the above experiment. The experiment time is less (1.2), but it is greater than the above experiment. Correctly Classified Instances of the experiment are almost equal to the above experiment, which is 88.10%. and Incorrectly Classified Instances are also equal which is 11.89%. TP Rate of the experiment is similar to the above experiment (0.88%). FP Rate is less than the previous one which is (0.37%). The precision of the experiment is 0.87% and less than experiment 4. The recall is large (0.88%). ROC Area is less than from all the above experiments (0.80%).

Scenario #6General decision tree with pruning with all attribute

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.957	0.466	0.91	0.957	0.933	0.845	low

	0.534	0.043	0.718	0.534	0.613	0.845	normal
Weighted Avg.	0.886	0.394	0.877	0.886	0.879	0.845	

=== Confusion Matrix ===

a b <-- classified as

3354 150 | a = low

333 382 | b = normal

Result analysis of scenario #6

The tree is complicated and not visible, leaves (1093) and the tree size is (1346). The computation time is 0.19 second which is less than the above experiment. CCI is high (88.55%) and ICI is 11.45% which is a little smaller than the above experiment. TP Rate is 0.87% which is less than the previous one. FP Rate is greater than that of the previous which 0.40% is. The precision of the experiment is 0.88% which is high and larger than experiment 5. The recall of the experiment is the larger from all the experiments done previously which is 0.89%. The ROC Area is 0.85%; it is larger than the previous experiment.

Scenario #7Binary decision tree with pruning with reduced attributes

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.958	0.503	0.903	0.958	0.93	0.844	low
	0.497	0.042	0.706	0.497	0.583	0.844	normal
Weighted Avg.	0.88	0.425	0.87	0.88	0.871	0.844	

=== Confusion Matrix ===

a b <-- classified as

3356 148 | a = low

360 355 | b = normal

Result analysis of scenario #7

The decision tree is the least complicated from all the experiments with a leaves of 321 and the tree size of 641. The experiment time is 1 second which is larger than the previous experiment. CCI of the experiment is 87.96% and it is less than the previous experiment. The ICI is larger than the previous experiment which is 12.04%. TP Rate of the experiment is 0.88% which is larger than the above experiment. And FP Rate is the larger one from all the experiments done previously which is 0.43%. The precision is 0.87% and smaller than the above experiment. The recall of the experiment is 0.88% which is a little smaller than the above experiment. ROC Area is high and enough to be accurate which is 0.84%.

Scenario #8General decision tree with pruning with reduced attributes

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.975	0.541	0.898	0.975	0.935	0.85	low
	0.459	0.025	0.792	0.459	0.581	0.85	normal
Weighted Avg.	0.888	0.454	0.88	0.888	0.875	0.85	

=== Confusion Matrix ===

a	b	<-- classified as
3418	86	a = low
387	328	b = normal

Result analysis of scenario #8

The decision tree is more complicated than experiment 7, with 760 leaves and 916 tree size. The time of computation is: 0.13 which is less than experiment 7. CCI of the experiment is 88.78% which is the larger one of all the experiments and the ICI is the smallest of all the experiments done. TP Rate of the experiment is 0.89% which is higher than all the above experiments. The FP Rate is 0.45% and larger than the above experiment. The precision is 0.88% which is high and larger than the above experiment. The recall of the experiment is larger than all the above experiments done which is

0.89%. The ROC Area is high and larger than experiment 7 which is 0.85% and it is possible to say the model is accurate.

In general the performance comparison of 8 experiments was presented. The best model done by confidence factor 0.5 and has a leaves of 760 and tree size of 916. The computation time for the model is 0.13. The CCI is 88.78% and ICI is 11.21%. TP Rate of the model is 0.89% and FP Rate 0.45%. The precision of the model is 0.88%. The recall is 0.89% and ROC Area is 0.85%.

The model has good precision and accuracy. Also the researcher tried to increase it into a better one but due to the reason of the data set which is not uniformly distributed (there is imbalance class) and the replacement of missing values the trial is not success full.

As it can be seen in the general summary of experiment table, the model has the best performance during experiment # 8 with reduced attributes compared to other experiments. The general summary of the experiments shows in the table 5.9, all the values of the experiment for comparing the best model.

Table 4.4 General summary of experiments

Experiment #	Type	Pruned	Leaves	Tree size	Time	CCI in %	ICI in %	TP Rate	FP Rate	Precision	Recall	ROC Area
1	Binary	No	649	1297	0.42	87.72	12.28	0.88	0.35	0.87	0.88	0.84
2	General	No	1797	2225	0.19	88.05	11.95	0.88	0.32	0.88	0.88	0.86
3	Binary	No	423	845	0.36	87.82	12.18	0.88	0.41	0.87	0.88	0.85
4	General	No	1264	1513	0.06	88.40	11.59	0.88	0.41	0.88	0.88	0.86
5	Binary	Yes	479	957	1.2	88.10	11.89	0.88	0.37	0.87	0.88	0.80
6	General	Yes	1093	1346	0.19	88.55	11.45	0.87	0.40	0.88	0.89	0.85
7	Binary	Yes	321	641	1	87.96	12.04	0.88	0.43	0.87	0.88	0.84
8	**General**	**Yes**	**760**	**916**	**0.13**	**88.78**	**11.21**	**0.89**	**0.45**	**0.88**	**0.89**	**0.85**

4.5 Generating rules from decision tree

In section 5.2 Experiment 8 has been selected for the best model of the dataset. The decision tree of the model shows the rules of the model by traversing from the root node till the leaf. Below rules extracted from the tree and selected by the researcher as most important to predict CD4 status of the patient following ART is shown.

Rules extracted from J48 decision tree

- When the patient is eligible to ART because of the CD4 count and if the patient started ART in the year 2003, the CD4 status of the patient for the next visit will be expected to Low.
- When the patient is eligible to ART because of the CD4 count and if the patient started ART in the year 2004 that the marital status of the patient is one of Married, Divorced, Widow and Separate, the patient will develop a normal CD4 status.
- When the patient is eligible to ART because of the CD4 count and if the patient started ART in the year 2005, and the OA weight of the patient is between 3K.g to 52.5 K.g, the CD status of the patient for the next visit will be Low.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and also the OA weight of the patient is between53K.g to 77 K.g and the OA WHO stage of the patient is stage 2 and if the patient is taking “1a40”,”1c” regimen then the patient will develop normal CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and the OAWHO stage of the patient is stage4 and the marital status is one of never married, Divorce, separate will probably develop low CD4 status. But if the patient is widow and current regimen the patient taking is “1C” will develop a normal CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and the OA weight of the patient is between 77.5K.g to 102K.g, then the patient will highly develop a low CD4 status.

- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2006 and the OAweight of the patient is 3K.g to 27.5K.g and the regimen taking is “1a30” the patient will develop low CD4 status for the next visit.
- If the eligible reason of the patient is CD4 and ART start year of the patient is 2006 and OAweight of the patient is between 28K.g to 52.5K.g, then the patient will probably develop low CD4 status for the next visit.
- When the eligible reason of the patient for the treatment is CD4 and if the patient started ART in the year 2007 and the OAweight is between 3K.g to 27.5K.g, then for almost all regimens the patient will develop low CD4 status.
- When the patient is eligible due to the reasons Clinical and TLC and current regimen that the patient is taking “1a30” and if the patient didn’t take any past ARV treatment will develop a normal CD4 status for the next visit but for those patients who did take past ARV treatment will be expected low CD4 status for all ART start years.
- When the patient is eligible due to the reasons Clinical and TLC and the WHO stage of the patient is stage 1 and if the patient did not take past ARV treatment, then the patient will develop normal CD4 status but if the patient took treatment the CD4 status is expected to be Low.
- When the patient is eligible due to TI and the OAweight of the patient is in between 3K.g to 52.5K.g then the patient will develop a Low CD4 status for the next visit.
- When the patient is eligible because of clinical diagnosis only and if the patient is registered for ART in the year 2003 and 2004 and if the patient did not kook any family planning methods it is expected to develop a low CD4 status for the next visit.
- When the patient is eligible because of clinical diagnosis only and if the patient is registered for ART in the year 2006 and if the regimen the patient is taking is “1a30” and if the patient did not took past ARV treatment then a normal CD4 status will be expected.

- When the patient is eligible because of clinical diagnosis only and if the patient is registered for ART in the year 2006 and if the regimen the patient is taking is "1b30" and if the patient OAweight is in between 53K.g to 77K.g is expected to develop low CD4 status.
- When the patient is eligible because of clinical diagnosis only and the ART started year is 2008 and marital status is never married, the patient will develop a normal CD4 status for the next visit.
- When the patient is eligible to ART because of clinical diagnosis and the patient started ART in the year 2009 and the regimen that the patient is taking is "1a30" and "1c" then will develop a normal CD4 status and almost all the rest regimens develop low CD4 status.
- If the patient is eligible because of Clinical diagnosis and the ART start year is 2010, then the patient will develop a low CD4 status for the next visit.
- If the patient is eligible because of Clinical diagnosis and the ART start year is 2011 and the marital status of the patient is one of married, never married and widow ,then the patient will develop a normal CD4 status for the next visit.

4.6 Rule generating using PART rule induction Algorithm

As discussed in the methodology section 1.4.2.5 PART is the second method in to the study for generating rule and makes comparison of the two results. PART rule induction algorithm is used after building models using J48 decision tree. The model selected from J48 decision tree is compared with model of PART with the same values of parameters of confidence factor, pruning, and attribute number.

PART rule induction apply an iterative process of consisting first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set, to remain with no examples left to cover the process repeatedly iterate.(Witten and Frank,2005)

The PART rule induction algorithm experiment is done with reduced 10 attributes, and a confidence factor of 0.5. The PART rule induction algorithm generates rules on a plain text format that is very easy to understand the generated rules which is the form of if then. The experiment summary is shown bellow.

=== Run information ===

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

Relation: ART1-weka.filters.unsupervised.attribute.Remove-R1-3,11-weka.filters.unsupervised.attribute.Remove-R2

Instances: 8438

Attributes: 11

MaritalStatus

ARTStatus

FunctionalStatus

EligibleReason

ARTStartYear

FamilyPlanning

OAWeight

OAWHOstage

CurrentRegimen

PastARV

OACD4

Test mode: split 50.0% train, remainder test

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

Number of Rules : 267

Time taken to build model: 0.44 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	3690	87.4615 %
Incorrectly Classified Instances	529	12.5385 %
Kappa statistic	0.4836	
Mean absolute error	0.1608	
Root mean squared error	0.3118	
Relative absolute error	56.4541 %	
Root relative squared error	83.1131 %	
Total Number of Instances	4219	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.96	0.543	0.897	0.96	0.927	0.874	low
	0.457	0.04	0.699	0.457	0.553	0.874	normal
Weighted Avg.	0.875	0.458	0.863	0.875	0.864	0.874	

=== Confusion Matrix ===

a	b	<-- classified as
3363	141	a = low
388	327	b = normal

The result of the model of PART rule induction algorithm show that it has 259 rules generated, the time required for the computation is 0.44 second, CCI is 88.58% and ICI are 11.42%, TP Rate is 0.89% and FP Rate is 0.42%. The recall is 0.87% and ROC area is 0.89% it is enough to say the model is accurate. Some of the rules produced by PART rule induction algorithm and selected by the researcher are listed below:

- When the eligible reason of the patient is CD4 and the ART started year is 2006 and the OAWeight of the patient is between 28K.g to52.5K.g will be expected a low CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2005 and the OA weight of the patient is between 77.5K.g to 102K.g, then the patient will highly develop a low CD4 status.
- When the patient is eligible to ART because of the CD4 count and the patient started ART in the year 2006 and the OAweight of the patient is 3K.g to 27.5K.g and the regimen taking is "1a30" the patient will develop low CD4 status for the next visit.
- If the eligible reason of the patient is CD4 and ART start year of the patient is 2006 and OAweight of the patient is between 28K.g to 52.5K.g, then the patient will probably develop low CD4 status for the next visit.
- When the eligible reason of the patient for the treatment is CD4 and if the patient started ART in the year 2007 and the OAweight is between 3K.g to 27.5K.g, then for almost all regimens the patient will develop low CD4 status.
- When the patient is eligible due to the reasons Clinical and TLC and current regimen that the patient is taking "1a30" and if the patient didn't take any past ARV treatment will develop a normal CD4 status for the next visit but for those patients who did take past ARV treatment will be expected low CD4 status for all ART start years.
- When the patient is eligible due to the reasons Clinical and TLC and the WHO stage of the patient is stage 1 and if the patient did not take past ARV treatment, then the patient will develop normal CD4 status but if the patient took treatment the CD4 status is expected to be Low.
- When the patient is eligible due to TI and the OAweight of the patient is in between 3K.g to 52.5K.g then the patient will develop a Low CD4 status for the next visit.
- When the current regimen the patient is taking "1a30" and the ART start year of the patient is 2006 and if the patient did not took past ARV treatment, then low CD4 status is expected to develop.

To compare the best model from J48 decision tree algorithm and PART rule induction algorithm the following summary table displays the values of parameters used for comparison.

Table 4.5 Summary of J48 and PART

Performance measure	J48 decision tree	PART rule induction
Precision	0.88	0.86
CCI	88.78	87.45
Tree size/number of rules	916	267

As it can be seen from the summary table of J48 and PART precision of J48 is a little bit larger than PART. And also CCI of J48 is larger than PART's. Even though the precision and CCI of the J48 is larger to PART, the size of rules produced by PART is more manageable and readable than J48. Based on the comparison criteria J48 decision tree is better than part rule induction algorithm. Since J48 is better than PART the model is created by J48. The created model produced rules and the from the rules the researcher fund that, Eligible Reason, ART Start Year, OA Weight, OA WHO stage, Marital Status, Current Regimen, Family Planning, Past ARV, Family Planning the most deterring factors that can predict the CD4 status of patients following ART.

CHAPTER FIVE

Conclusion and Recommendations

6.1 Conclusion

Since then, HIV/AIDS has become a major public health concern in the country, leading the Government of Ethiopia to declare a public health emergency in 2002. In 2007, the estimated adult HIV/AIDS prevalence in Ethiopia was 2.1 percent. Although the epidemic is currently stable, HIV/AIDS remains a major development challenge for Ethiopia. Poverty, food shortages, and other socio-economic factors amplify the impact of the epidemic.

In the last few years, the Government of Ethiopia has increased efforts to accelerate progress toward universal access to HIV prevention, treatment, care, and support, emphasizing involvement of local stakeholders and decentralization. The Government trained 32,000 health extension workers to promote community health initiatives, aid households, and deliver ART as part of the decentralization.

Antiretroviral Therapy (ART) is treatment for AIDS that helps the body's immune system recover from the damage caused by infection with HIV. Although ART cannot cure AIDS, persons on ART will begin to feel better, eat more, and put on weight. Their bodies will recover the ability to fight infections. They recover their sense of hope for the future and can become powerful advocates for prevention and mitigation of HIV in their families and communities. They may remain well for many years, but must continue to take Antiretroviral (ARVs) for the rest of their lives. Thus, ART is an important component of the global response to AIDS.

HIV most often infects CD4 cells. The genetic code of the virus becomes part of the cells. When CD4 cells multiply to fight an infection, they make more copies of HIV. When someone is infected with HIV but has not started treatment, the number of CD4 cells they have goes down. This is a sign that the immune system is being weakened. The lower the

CD4 cell count, the more likely the person will get sick. There are millions of different families of CD4 cells. Each family is designed to fight a specific type of germ.

The study data was taken from two hospitals of the south west of Ethiopia namely Jimma and Bonga hospitals. The study followed the CRISP-Dm data mining methodology, it do have six phases called: Business Understanding, Data Understanding, Data preparation, Model building, Evaluation and deployment. The study used a total of 8438 records which is a sum of 6242 from Jimma hospital and 2197 from Bonga hospital. After the records obtained some measures has taken. In the phase of data preparation Data transformation, data discritazation and missing value handling using the modal and mean value replacement has done.

The format of Jimma hospital database is SQL server database management (DBMS) and Bonga hospital uses Microsoft Access at the back end. Both databases have a front end made from Visual Basic to enter the data. Systool recovery software has been used to copy the records of SQL data of the Jimma ART data. Microsoft Excel used for combining the two hospitals records and to convert the record into Comma Separated Values (CSV) format.

From known tasks of Data mining, the study selected classification method to predict the status of CD4 of patients following ART. Since J48 is a technique for classification and it is reliable and efficient technique for providing high classification accuracy for a simple representation of gathered knowledge the study applied it as one technique to build the model. And PART rule induction algorithm is also important to obtain easily interpretable rules; the study selects this algorithm to use it in parallel with J48 decision tree algorithm.

Confidence factor has impact in the classification of instances into correctly and incorrectly one. When the researcher conducted different experiments by changing confidence factors in the increasing order, the incorrectly classified instances decrease the same time up to it reach a confidence factor of 0.5. Since it is good to use minimum incorrectly classified instances, confidence factor 0.5 is the best to to build the model.

The attributes ranked for predicting CD4 status of patients following ART are 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 best performance achieved by J48 decision tree algorithm is a generalized decision tree with pruning with reduced attributes. The model classifies instances correctly 88.79% and incorrectly classifies 11.21%. The weighted average precision of the model is 0.88 with recall of 0.89 and ROC area of 0.85. The model has 760 numbers of leaves and 916 size of tree. The time taken to build the model is 0.05 seconds. The analysis of this model shows that the model is quit efficient to predict CD4 status of patients following ART.

Finally in order to conclude the classification done using J48 decision tree is the best model than PART rule induction algorithm. J48 algorithm is effective to predict the CD4 status of patients following ART. From the model built it is fund that attributes: Eligible reason, ART status, ART start year, OAweight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status, Past ARV are the most determining factors for predicting CD4 status.

6.3 Recommendations

- Health facilities record keepers especially ART data clerks need to give attention in the recording of patient's information; this is because the records are important to predict the CD4 status of patients.

- Ministry of Health should work on a national standard data base format for recording ART data, because if the data bases are not in the same standard it is difficult for researchers to combine different health facilities record.

- Ministry of Health needs to give attention on the use of Data mining on the electronics Health records especially HIV/AIDS records.

- Since the patterns obtained are important to deploy and uses for decision support, the researcher leave open for the others investigators to work on the deployment of the model for Hospitals.

- Health facilities need to improve the accessibility of data for the researchers.

References

ETC (AIDS Education and Training Center). 2011.aidsinfonet.Fact Sheets Vol.124 P-1.

Bramer.M. 2007. Principles of Data mining. Springer. London limit.

David. H 2001.Principles of Data mining. Massachusetts London, England. MIT press.

Fayyad, Piatesky-Shapiro.G and Smyth.P. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework.CA.

Gadelha.A. 2002. Morbidity and survival in advanced AIDS in RiodeJaneiro. Revised Ed. TropPaulo.

Han J, Kamber M. , 2006 Data Mining: Concepts and Techniques. Morgan Kaufmann. Second edition.

Hristidis.V.2010. Information Discovery on Electronic Health Records. U.S.A Chapman &Hall/CRC Taylor & Francis Group .

Kantardzic.M.2003. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Inc.

Larose, Daniel T., 2005. Discovering knowledge in data : an introduction to data mining. John Wiley & Sons, Inc., Hoboken, New Jersey.

Makie T, et al.2010. Estimating $CD4^{+}$ Cell Counts of an Individual using Population Historical Data. J AIDS Clinic Res 1:104.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Marc M., VAN. H .Data mining. Katholieke Universiteit Leuven press.

Mellors. J, Munoz. A .1997.Plasma viral load and CD4+lymphocytes As prognostic markers of HIV-1 infection. Ann Intern Med.

MOH/HAPCO. 2006. AIDS in Ethiopia Technical documents for the sixth reports. Addis Ababa, Ethiopia.

Osmar R. Zaïane, 1999. Principles of Knowledge Discovery in Databases. University of Alberta press.

Pathfinder International. 2007. The Essentials of Antiretroviral Therapy for Health care and program managers. Technical Guidance series no.5 U.S.A

Sumathi.S,. Sivanandam. S.N.2006. Introduction to Data Mining and its Applications. Springer-Verlag Berlin Heidelberg.

Two Crows Corporation and Training center. 1999. Introduction to Data mining and Knowledge Discovery. U.S.A

Two Crows Corporation and Training center. 2005. Introduction to Data mining and Knowledge Discovery. 3rd Edition. U.S.A

Witten IH, Frank E. 2005 Data Mining: Practical Machine Learning Tools and Techniques. 2nd edition. Morgan Kaufmann.

XindongWu and Vipin Kumar. 2008. The top ten Algorithms in data mining. Chap man and hall/CRC press Tayler and fransis group.USA

USAID/ETHIOPIA. 2010. HIV/AIDS Health profile.

WHO. 2006. Antiretroviral Therapy for HIV infection in adult's adolescents. 2nd Edition.

ANNEX

Sample CSV format of the data set

Sex	Age	ReligionID	MaritalSta	Education	ARTStatus	Functiona	EligibleRe	ARTStartY	FamilyPla	Pregnant	OAWeight	OAWHOst	CurrentRe	PastA
Female	15-24	Orthodox	Married	Primary	OA	W	ClinicalTL	2011	No	No	3-27.5	Stage3	1c	Yes
Male	25-49	Orthodox	Widow	Noeducat	OA	W	Clinical	2008	No	No	3-27.5	Stage3	1a30	Yes
Female	15-24	Orthodox	Married	Primary	OA	W	Clinical	2011	No	No	3-27.5	Stage3	4a	No
Female	25-49	Orthodox	Married	Primary	OA	W	Clinical	2011	No	No	3-27.5	Stage3	4a	No
Female	25-49	Orthodox	Married	Primary	OA	W	Clinical	2011	No	No	3-27.5	Stage3	4a	No
Female	50-64	Orthodox	Married	Primary	OA	W	Clinical	2011	No	No	3-27.5	Stage3	4a	No
Male	25-49	Orthodox	Married	Primary	OA	W	CD4	2007	No	No	3-27.5	Stage2	4c	No
Female	25-49	Orthodox	Married	Primary	OA	W	CD4	2007	No	No	3-27.5	Stage2	4c	No
Female	25-49	Orthodox	Married	Primary	OA	W	CD4	2007	No	No	3-27.5	Stage2	4c	No
Female	25-49	Orthodox	Nevermar	Primary	OA	W	Clinical	2010	No	No	3-27.5	Stage2	1c	Yes
Male	15-24	Muslim	Married	Primary	OA	W	TLC	2005	Yes	No	28-52.5	Stage4	1a40	No
Female	25-49	Muslim	Married	Primary	OA	W	TLC	2005	Yes	No	28-52.5	Stage4	1a40	No
Female	25-49	Muslim	Married	Primary	OA	W	TLC	2005	Yes	No	28-52.5	Stage4	1a40	No
Male	15-24	Muslim	Married	Secondary	OA	W	TI	2006	Yes	No	53-77	Stage3	1c	No
Female	15-24	Orthodox	Married	Primary	OA	W	CD4	2006	No	No	3-27.5	Stage4	4c	No
Female	25-49	Orthodox	Married	Primary	OA	W	CD4	2006	No	No	3-27.5	Stage4	4c	No
Female	0-14	Orthodox	Married	Primary	OA	W	CD4	2008	No	No	3-27.5	Stage1	4a	No
Female	15-24	Orthodox	Married	Primary	OA	W	TI	2007	No	No	3-27.5	Stage3	4c	No
Female	15-24	Orthodox	Married	Primary	OA	W	TI	2007	No	No	3-27.5	Stage3	4c	No
Male	50-64	Orthodox	Married	Primary	OA	W	TI	2007	No	No	3-27.5	Stage3	4c	No
Male	50-64	Orthodox	Married	Primary	OA	W	TI	2007	No	No	3-27.5	Stage3	4c	No

Sample result of Experiment#1

```
EligibleReason = Clinical
|   ARTStartYear = 2003: low (19.0)
|   ARTStartYear != 2003
|   |   ARTStartYear = 2004
|   |   |   MaritalStatus = Widow: normal (4.0)
|   |   |   MaritalStatus != Widow
|   |   |   |   FamilyPlanning = No: low (38.0)
|   |   |   |   FamilyPlanning != No
|   |   |   |   |   MaritalStatus = Married
|   |   |   |   |   |   OAWeight = 28-52.5: low (3.0)
|   |   |   |   |   |   OAWeight != 28-52.5: normal (3.0)
|   |   |   |   |   MaritalStatus != Married: low (3.0)
|   |   ARTStartYear != 2004
|   |   |   ARTStartYear = 2005
|   |   |   |   OAWHOstage = Stage2: normal (4.0)
|   |   |   |   OAWHOstage != Stage2
|   |   |   |   |   OAWHOstage = Stage1: low (7.0)
|   |   |   |   |   OAWHOstage != Stage1
|   |   |   |   |   |   Age = 50-64: low (6.0)
|   |   |   |   |   |   Age != 50-64
|   |   |   |   |   |   |   FunctionalStatus = B: low (5.0)
|   |   |   |   |   |   |   FunctionalStatus != B
|   |   |   |   |   |   |   |   Age = 0-14: low (4.0)
|   |   |   |   |   |   |   |   Age != 0-14
|   |   |   |   |   |   |   |   |   CurrentRegimen = 1a40
|   |   |   |   |   |   |   |   |   |   EducationalLevel = Secondary
|   |   |   |   |   |   |   |   |   |   |   Religion = Muslim: normal (2.0)
```

Sample result of Experiment#2

```
EligibleReason = CD4
|   ARTStatus = IN: low (1582.0)
|   ARTStatus = OA
|   |   ARTStartYear = 2003: low (14.0)
|   |   ARTStartYear = 2004
|   |   |   MaritalStatus = Married: low (58.0/3.0)
|   |   |   MaritalStatus = Nevermarried
|   |   |   |   Religion = Orthodox: normal (3.0)
|   |   |   |   Religion = Protestant: low (0.0)
|   |   |   |   Religion = Muslim: low (3.0)
|   |   |   |   Religion = Catholic: low (0.0)
|   |   |   |   Religion = Other: low (0.0)
|   |   |   MaritalStatus = Divorced: low (0.0)
|   |   |   MaritalStatus = Widow: low (3.0)
|   |   |   MaritalStatus = Separated: low (0.0)
|   |   |   MaritalStatus = W: low (0.0)
|   |   ARTStartYear = 2005
|   |   |   OAWeight = 3-27.5: low (14.0)
|   |   |   OAWeight = 28-52.5
|   |   |   |   CurrentRegimen = 1a30: low (225.0)
|   |   |   |   CurrentRegimen = 4a: low (0.0)
|   |   |   |   CurrentRegimen = 4b: low (0.0)
|   |   |   |   CurrentRegimen = 4c: low (4.0)
|   |   |   |   CurrentRegimen = 4d: low (0.0)
|   |   |   |   CurrentRegimen = 1a40: low (45.0)
|   |   |   |   CurrentRegimen = 1b30
|   |   |   |   |   FamilyPlanning = No
```

TABLE OF CONTENTS

LIST OF ABREVATIONS

AI	Artificial Intelligence
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
CAD	Computer Assisted Design
CCI	Correctly Classified Instances
CD4	Clustered Differentiation 4
CRISP-DM	Cross Industry Standard for Data Mining
CSV	Comma separated Value
DM	Data Mining
EHR	Electronic Health Record
EIS	Executive Information System
FPR	False Positive Rate
FuReA	An adaptive fuzzy regression technique
HAART	Highly Active Antiretroviral Therapy
HAPCO	HIV/AIDS Prevention and Control Office
HIV	Human Immunodeficiency Virus
ICI	Incorrectly Classified Instances
IDU	Injection Drug Use
IRM	International Resource Management
KDD	Knowledge Discovery in Database
MAC	Millennium AIDS Campaign
MOH	Ministry of Health
NAC	National Aids Counsel
OA	On ART
RDBMS	Relational Database Management
ROC	Receiver Operating Characteristics
TDIDT	Top Down Induction of Decision Tree
TLC	Total lymphocyte count
TPR	True Positive Rate
UNAIDS	Joint United Program Nations Program
VCT	Voluntary Counseling and Testing
VRML	Virtual Machine Language
WHO	World Health Organization
WWW	World Wide Web

LIST OF TABLES

LIST OF FIGURES