JIMMA UNIVERSITY

COLLEG OF NATURAL SCIENCES

DEPARTMENT OF INFORMATION SCIENCE

PREDICTING SURVIVAL RATE OF HIV/AIDS PATIENTS: THE CASE OF JIMMA UNIVERSITY SPECIALIZED HOSPITAL

BY

YETNAYET TADESSE

JIMMA, ETHIOPIA

June, 2014

JIMMA UNIVERSITY

COLLEG OF NATURAL SCIENCES

DEPARTMENT OF INFORMATION SCIENCE

PREDICTING SURVIVAL RATE OF HIV/AIDS PATIENTS: THE CASE OF JIMMA UNIVERSITY SPECIALIZED HOSPITAL

BY

YETNAYET TADESSE

A THESIS SUBMITTED TO THE DEPARTMENT OF INFORMATION SCIENCE OF JIMMA UNIVERSITY IN PARTIAL FULFILMENTS FOR THE DGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE (IKM).

Principal Advisor: Rahel Bekele (Dr.)

Co-Advisor: Mr. Samuel Sisay (Msc.)

JIMMA, ETHIOPIA

June, 2014

JIMMA UNIVERSITY

COLLEG OF NATURAL SCIENCES

DEPARTMENT OF INFORMATION SCIENCE

PREDICTING SURVIVAL RATE OF HIV/AIDS PATIENTS: THE CASE OF JIMMA UNIVERSITY SPECIALIZED HOSPITAL

BY

YETNAYET TADESSE

Members of the examining board:

Name	Title	Signature	Date
_____	Chair person	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

DECLARATION

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

_____

Yetnayet Tadesse

June, 2014

This thesis has been submitted to the department for examination with our approval as university advisors:

Principal Advisor: Rahel Bekele (Dr.)…….

Co-Advisor: Samuel Sisay (MSc.)…….

June, 2014

Dedicated to

My mother: Yetmwork Tafesse

&

My father: Tadesse Bedore

ACKNOWLEDGMENT

I would like to acknowledge my advisor Dr. Rahel Bekele for her constructive comments, great deal of patience and guidance that she has provided me throughout the study.

My deepest gratitude also goes to my co-Advisor Mr. Samuel Sisay (Samiy), for his support, encouragement and comments on every aspect of the problem I faced during the course of this work.

I would also like to thank Jimma University, Department of Information Science for financial support and overall facilitation of the research from the beginning until the end.

I thank the staff in Jimma University hospital ART clinic especially data clerk Ayantu. I couldn't have started the study, if it was not for the interest and support contributed from her.

My special thanks goes to my Friends Chakich, mamush and Jossi for the constant assistance and the times we have had during my study from elementary to date.

I would like to thank my classmates specially Mesfin Alemu, Chala Diriba and Tilahun Sheferaw for their comments, constructive ideas, advices and suggestions.

Finally I am also grateful to many others, who helped in one way or another during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACCRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ARFF	Attribute Relation File Format
ART	Antiretroviral Therapy
ARV	Antiretroviral
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma Separated Value
FHAPCO	Federal HIV/AIDS Prevention and Control Office
FMOH	Federal Ministry of Health
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunity Virus
ICAP	International Center for AIDS Care and Treatment Programs
IQR	Inter Quartile Range
OART	On Antiretroviral Therapy
PART	Partial Decision Tree
ROC	Receiver Operating Characteristics
SEMMA	Sample Explore Modify Model Asses
SPSS	Statistical Package for Social Science
SQL	Structured Query Language
UNAIDS	Join United Nations program on HIV/AIDS
WEKA	Waikato Environment for Knowledge Learning
WHO	World Health Organization

ABSTRACT

Recent reports from WHO and UNAIDS indicate that the number of people using ART are increasing from time to time. This number is dramatically increasing in sub Saharan African countries including Ethiopia. The introduction of ART has greatly improved the survival of HIV/AIDS infected people. However, the survival rate of these ART users needs further investigation so as to demonstrate what could be done for further improvement. The main objective of this study was, therefore, to develop a predictive model for the survival rate of HIV/AIDS patients. The overall activity of this thesis is guided by a Hybrid-DM model which is a six step knowledge discovery process model. The study has used 7,700 instances, twelve predicting and one outcome variables to run the experiments. The mining algorithms; J48 and PART are used in all experiments due to their popularity in recent related works. Ten-fold cross validation and 70/30 split criteria test option were used to train and test the classifier models. Performance of the models is compared using accuracy, TPR, FPR, and the area under the ROC curve. J48 algorithms were better performance with 97.4 % accuracy running on 10 cross fold test option with default parameter using 13 attribute than any experimentation done for this research purpose. The selected significant attributes of HIV/AIDS patients for the survival rate has been identified. These are Education Level, weight of patients, marital status, Baseline CD4 count, Eligibility reason, Drug regimen, Baseline WHO stage, age. This study identified attributes that significantly indicate the level of survival rate. Identifying patients at a low probability of survival has the advantage that due attention would be given to the risk group during their follow up to maximize the survival while they are taking ART.

Future works can be done on the degree of association that coexists between those attributes and survival rate is forwarded.

CHAPTER ONE

1. INTRODUCTION

1.1.Background

There are many kinds of tropical infectious diseases in the world that affect human beings in different aspects. Tropical infectious diseases are diseases that thrive in hot, humid conditions, and affect human beings. Infectious diseases are caused by pathogenic microorganisms, such as bacteria, viruses, parasites or fungi; the diseases can be spread, directly or indirectly, from one person to another. One major tropical infectious disease that occurs in the tropics is HIV/AIDS. HIV/AIDS is a viral disease. A decade ago, being infected with AIDS was almost equivalent to a death sentence. Since 1996, with the introduction of combined antiretroviral treatment, AIDS has become chronic, but manageable disease (Steven et.al, 2003).

HIV/AIDS (Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome) is the most destructive disease and the world suffers a lot because of it. It is incurable and has no vaccine to its infection. Ethiopia is one of the Sub-Saharan African countries most severely affected by the HIV/AIDS pandemic. Currently, the national adult prevalence rate is estimated at 2.3 percent and an estimated number of 1.2 million people are living with HIV/AIDS (UNAIDS/WHO, 2010).

Antiretroviral Therapy (ART) is a treatment for people who are infected with human immunodeficiency virus using anti-HIV drug (WHO, 2012). ART means treating retroviral infections like HIV with drugs. The drugs do not kill the virus (Antiretroviral Therapy Cohort Collaboration, 2008.), however they slow down the growth of the virus.

ART is treatment of people infected with human immunodeficiency virus using anti-HIV drug (WHO, 2012). The standard treatment consists of a combination of at least three drugs (often called HAART) that suppress HIV replication (WHO, 2012).

The introduction of ART has greatly improved the survival of HIV/AIDS infected people. ART reduces morbidity and mortality by suppression of viral replication, restoration and preservation of immune function, and prevention of drug resistance. The treatment keeps the patient alive for unknown time. A valid estimation of life expectancy and knowing the significant attribute for survival after diagnosis would be of great value. The values of predicting the survival rate of HIV/AIDS patient is about prediction of the future course and giving hope. This is also good hope for health care policy makers and HIV/AIDS control programs. At the individual level the predicted information can help patients make informed decisions with regards to the quality of life.

1.2.Statement of the Problem

Though the burden of infectious disease has diminished globally, infectious diseases are still a major obstacle to development in Africa (African union, 2013). These diseases continue to perpetuate inequality and poor health, and development outcomes, especially for the marginalized rural and urban unfortunate people. It is a well known truth that Ethiopia is one of the most countries smack by the HIV/AIDS epidemic. The epidemic has claimed the lives of the country's adult that would otherwise have contributed immensely to the country's development in many aspects. Ever since the first infections were identified in the country, a lot has been done

both by the government and other stake holders to curb the effect of the disease specially to suppress the risk of transmission.

Numerous researches have also been conducted that tried to tackle many of the issues that arise in connection with the HIV epidemic. One thing that should, however, to be noted is that many of these research works mainly focused on the assessment of the prevalence and the study of the numerous prevention measures that should be undertaken to stop or reduce the spread of the epidemic. Little attention seems to have been given to study the situation of those living with the virus. Especially when seen in light of the current situation in the country where we have over 1 million individuals living with the virus and over 1 billion birr allocated for HIV mobilization in 2008/2009 only (FHPCO, 2010), it is important that research works focus on the study of people living with HIV/AIDS (PLWHA) situation. This is more so when we see the large number of HIV positive people who are ON ART.

Thus, it should be stated at this stage that researches should be conducted to evaluate the effectiveness of the free antiretroviral drugs treatment option. Though research work is not non-existent in this area but the problem is that all those previous studies were conducted by using a very small proportion of the database. Besides, in those studies, data analysis was conducted by using simple statistical techniques (such as regression and verification techniques). Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world databases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases (Plate et. al., 1997). Many of the factors that can possibly affect the survival of people who are in ART follow up remain unstudied.

Given this as a back drop, many important questions related to the best use of ART remain unanswered, including for example, how long will treatment keep the patient alive? Which attributes are more important to predict the survival rate of ART following patients? This thesis is focused on the consideration of some of the possible attributes that may possibly influence the survival status of patients who are following ART in Jimma University specialized hospital. The dynamics of the restoration of attributes of survival under these circumstances are not well understood. The values of predicting the survival rate of HIV/AIDS patient is a prediction of the future course and outcome of HIV/AIDS. The information gained from the survival rate prediction were used to determine health care policies, to monitor the progress of HIV/AIDS control programs, as a tool to assess the efficiency of treatment protocols, to monitor the progress of treatment programs and as an aid in choosing treatment types and methodologies.

To come up with the best predictive model this research attempted to answer the following research questions:

1. Which attributes are more important to predict the survival rate of ART following patients?
2. Which classification algorithm is more suitable to build a survival rate predictive model for ART following patients?

1.3.Objectives of the Study

1.3.1. General objective

The main objective was to develop a predictive model for the survival rate of HIV/AIDS patients.

1.3.2. Specific Objectives

The specific objectives of the study were:

1. To identify the important attributes that help to predict the survival rate of ART following patients.
2. To find classification algorithm that will be more suitable to build a survival rate predictive model for ART following patients.
3. To apply classification algorithms to train, test and build classifier models.
4. To compare and suggest the best model for prediction of HIV/AIDS patients survival rate
5. To validating the proposed model

1.4.Significance of the study

I. The result of this study provides information to government and other stake-holders in setting policies, strategies, and further investigation for survival of HIV-infected patients.

II. This result will help both donors and government to appreciate factors that influence the survival status of HIV/AIDS patients and adjust their intervention programs accordingly.

III. The finding of this study will enable clinicians to target services for patients at exceptionally low survival rate and to provide patients with more accurate predictive information.

IV. The results will also contribute a lot for further study in the field.

1.5.Scope and Limitation of the Study

This study is limited to predicting survival rate of patients based on useful attributes found in ART records kept in databases of Jimma University specialized Hospital.

The first exclusion criterion is accompanied with the enrollment year of the patients in to the program. Therefore, the study dataset records comprise those which are only recorded from 1996 E.C to 2006 E.C. Thus, those which are found either above or below this two demarcation years were discarded from the study dataset.

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

Regarding the data mining tasks, classification mining techniques were used for the dependent variable. Among the available classification algorithms, decision tree (J48) and rule induction (PART) were used to build the models.

The study is limited to developing a predictive model from attributes found in the ART dataset. But these dataset are not enough to make prediction of survival rate of HIV/AIDS patient. There exist variables like feeding style, drug abuse, economical status, and others which might have a role in determining one's survival rate. Therefore, missing such variables within the study dataset has an effect on the accuracy of the findings.

1.6.Research Methodology

The overall activity of this thesis were used a Hybrid data mining model which is a six step knowledge discovery process model. Due to the nature of the problem and attributes in the dataset, classification mining task were selected to build the classifier models. The mining algorithms; J48 and PART were used in all experiments due to their popularity in recent related works.

Hybrid process model were selected as a better model regarding its design to suit for academic researches. Accordingly, to realize a model that yields optimum classifier of survival rate of an individual, Hybrid data mining process model were used to guide the overall execution of the project.

1.6.1. Source and types of data

Secondary data was used as a source of information. The secondary data included patient's data history from Jimma University specialized hospital.

1.6.2. Methods of Data Collection

Collecting representative subset of ART data was a prerequisite to address the objective of this research. Therefore, Data were collected from the records of HIV patients who were in follow up treatment in the ART clinic of Jimma university hospital ART Data bases. The Data base was in SQL server through which the data clerks can enter patient data and generate different reports. A full backup of the database of the ART was taken from Jimma university Hospital. It stores data on HIV/AIDS patients' who are in care to start drug, Eligible to start date and on those who have already started using ART drug. The dataset stores eleven years record from the year 1996 to 2006. The total dataset obtained was 9200.

1.7.Ethical Consideration

The ethical approval and clearance were obtained from Jimma University Research Ethical Board. The necessary explanation about the purpose of the study and about its procedure, assurance of confidentiality, the right not to participate on the study without any consequences was included.

1.8.Organization of the Thesis

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

The second chapter focuses on literature review on HIV/AIDS and ART and General overview of ART also extensive review of related works are included.

Chapter three is about Data mining concepts and techniques of this study. Data mining research literature review and related works done on HIV/AIDS and ART are included.

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

Chapter five is where the experiments conducted are presented. Here, topics about data mining, model selection and Prototype development were discussed in detail. Results of the experiments were also analyzed and interpreted.

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

CHAPTER TWO

2. HIV/AIDS AND ART

2.1. HIV/AIDS

The risk of HIV infection increases by number of sexual partners, intravenous drug use, any sex without condom, alcohol and other drug use, tattoos and body piercing with contaminated needles or instruments. Since AIDS was first identified, researchers and scientists have tried their best to find medicine and vaccine but the results everywhere were not successful.

Consequently the lives of human beings have been observed to be threatened due to infection with the virus. On top of all, since the disease is affecting the most productive citizens, it is natural that it causes damage to national economy (UNAIDS, 2000).

2.1.1. HIV/AIDS in Ethiopia

The first evidence of HIV epidemic in Ethiopia was detected in 1984. Since then, AIDS has claimed the lives of millions and has left behind hundreds of thousands of orphans (FHAPCO, 2007). Ethiopia is one of the hardest hit Sub-Saharan Africa countries by the HIV pandemic. In 2009 a single point estimate of AIDS related deaths for 2010 was 44,751 of which 7,214 were children and there were about 84,189 HIV positive pregnancies from which 14,140 gave birth to HIV positive children and this resulted in an estimated increase in the number of HIV positive children from 72,945 in 2009 to 79,871 in 2010 (National AIDS resource center, 2009).

According to the National Factsheet, 2010 of National AIDS resource center the total number of HIV positive people in 2010 is estimated to be 1,326,329 including 137,494 new HIV infections

and excluding 28,073 AIDS related deaths during the year. It is also estimated that a total of 90,610 HIV positive children under 15 years of age including 14,276 new infections and excluding 3,537 AIDS related deaths of children during the year.

According to the FHAPCO single point estimate for prevalence of HIV/AIDS in Ethiopia, the adult (15-49) HIV prevalence for 2007 is estimated at 2.1% of which 7.7% is urban and 0.9% is rural (EMOH, 2007). In 2010, the FHAPCO estimates of the overall adult (15-49) HIV prevalence is 2.4%. Urban and rural HIV prevalence rates were 7.7% and 0.9%, respectively. In 2010, an estimated 28,073 Ethiopians died of AIDS scaling the number of children who have lost one or both parents to AIDS to 804,184 (National Factsheet, 2010).

HIV/AIDS has been and still is the greatest challenges to the Ethiopian health system, as elsewhere in sub-Saharan African countries. It has remained among the major causes of deaths over the past two decades. In 2010, more than one million people are estimated to be living with HIV in Ethiopia of whom nearly 397,818 need ART care and treatment (National Factsheet, 2010).

2.1.2. Goals of ART

Infection with HIV causes immunologic deficiency that results in depletion of CD4 cells and suppression of cell mediated immune defenses. HIV infects CD4 cells by interacting with their CD4 receptors, which allows the virus to gain entry into the cells. The invading HIV replicates within the CD4 cells, destroys them and spreads to other CD4 cells, depleting the CD4 cell population. Because CD4 cells direct and activate immune responses, many immune functions degrade as a result of HIV infection (Virco, 2008). Individuals with weakened immune defenses

are susceptible to infections caused by opportunistic pathogens that do not normally cause disease for immune competent individuals. These infections are known as opportunistic infections. Once an opportunistic infection has begun, it can rapidly spread throughout the body via the circulatory system, damages vital organs and becomes fatal (Virco 2008).

The goals of therapy in treating HIV/AIDS infected individuals are, therefore, in view of the following points:

- Clinical goal which aims to extend life expectancy and quality of life for patients infected with HIV,
- Virological goal is to reduce the HIV viral load to the lowest level possible in order to prevent disease progression and limit development of resistance to ARV drugs,
- Immunological goal is to preserve and restore immunologic functioning in the normal range. This involves the quantitative component of CD4 cell count in the normal range. It also involves the qualitative goal of resisting infections by opportunistic pathogens, and
- Epidemiological goal is to reduce transmission of HIV to others. (Virco, 2008).

The clinical symptoms of HIV infection were evolved on the depletion of CD4 cells and the replication of HIV RNA. Therefore, CD4 cell count and HIV viral load are two of the most important predictors of the clinical prognosis of HIV-infected subjects.

Highly active antiretroviral therapy (HAART) and a single or combination of several drugs, have high activity to inhibit HIV RNA replication. HIV cocktail therapy is a combination reagents which inhibit the replication of HIV RNA at different stages of HIV life-cycle. The currently available HIV inhibition reagents can be categorized into Nucleotide reverse transcriptase

inhibitors, Non-Nucleotide reverse transcriptase inhibitors, protease inhibitors, and integrated zinc-finger inhibitors (Zhang, 2007).

2.1.3. Importance of ART in Ethiopia

In the face of competing demands such as malaria, TB, and famine some question whether an investment on ART in Ethiopia is justifiable. Given the impact of AIDS across society and the potential of ART to reduce the burden, the justification for pursuing this agenda is unarguable. HIV and AIDS are affecting every sector of Ethiopian society. At the macro level, the health, agriculture, education, business and industry sectors are all adversely impacted by the disease. Families and communities are likewise affected. The MOH estimate that the annual mortality rate for those in the 15-49 age range were increased from a projected 200,000 without factoring in AIDS to over 350,000 in 2004 (with the AIDS epidemic). Besides these numbers, the impact of AIDS increased absenteeism in the workplace, reduction of productivity, reduced family income, and increased family expenditure on health care and burial rituals. In a resource poor country such as Ethiopia, the economic impact of AIDS-related illness and death is severe (AIDS Resource Center et al, 2005).

In contrast to Ethiopia, AIDS-related deaths and illnesses in countries where ART has been available since the mid 1990s have considerably declined (UNAIDS, 2004). The experience of developed nations, as well as countries such as Brazil (from middle income countries which produces ART drugs), have proven that ART treatment reduces disease burden and dependence, and increases the function, well-being, and productivity of individuals. This in turn, can help offset some of the consequences of the HIV and AIDS pandemic (AIDS Resource Center et al, 2005).

2.1.4. General Overview of the Use of ART

Antiretroviral medications are designed to inhibit the reproduction of HIV in the body. The main effect of antiretroviral treatment is to suppress viral replication, allowing the individual's immune system to recover and protect him/her from the development of AIDS and death. In other words what this means is that if ART is effective, the deterioration of the immune system and the onset of AIDS can be delayed for years thereby improving the quality of life of the victims. Standard ART (also known as HAART) consists of the use of at least three ARV drugs to maximally suppress the HIV virus and stop the progression of HIV disease (WHO, 2012).

ART changes the natural history of HIV infection. Results from medical studies on HAART have been extremely impressive. Since its introduction in 1996, mortality and morbidity rates in HIV-infected individuals in countries with widespread access to HAART have plummeted. The use of antiretroviral medicines dramatically reduced AIDS related illnesses and death in countries where these drugs are widely accessible (4, 14). Although the treatments are not a cure and continue to present new challenges with respect to side-effects and drug resistance ART as disease modifying therapy for established HIV infection has produced dramatic effects on morbidity and mortality among HIV-infected patients. As a result of the widespread use of ART, the HIV/AIDS pandemic which was once regarded as an infectious disease with an almost universal fatal outcome has been transformed into a manageable chronic infectious disease (WHO, 2003)

A study in the US conducted from January 1994 through June 1997 which evaluated 1255 patients who were using antiretroviral revealed that mortality among the patients declined from 29.4 per 100 person-years in 1995 to 8.8 per 100 person-years in the second quarter of 1997. The

incidence of major opportunistic infections (OI) also declined from 21.9 per 100 person-years in 1994 to 3.7 per 100 person-years by mid-1997. There were also reductions in mortality and incidence of opportunistic infections regardless of sex, race, age, and risk factors for transmission of HIV (palella et.al, 2008).

In a study to evaluate changes and risk factors for death among HIV-infected children in Paediatric AIDS Clinical Trials Group 219/219c in US among 3553 HIV-infected children was followed up for a median of 5.3 years. The study shows that increased risk of death was significantly associated with low CD4, pneumonia and AIDS-defining illness at entry. Whereas, decreased risks of mortality were identified for children timely began highly active antiretroviral therapy (Brady et al., 2010). On the other hand a case control study conducted by Fontana and colleagues (1999), to see body composition in HIV infected children in relation with disease progression and survival using a total of 86 HIV infected and 113 HIV uninfected children showed that weight in HIV infected children was significantly less than in a control children with similar age; moreover weight were significantly associated with increased risk of death. Therefore the authors conclude that body weight for age is a good prognostic indicator (Fontana et al.1999). The department of pediatrics in New York University, New York, U.S.A. has conducted a study to measure the efficacy of highly active antiretroviral therapy in HIV infected children in resource poor setting. To measure the clinical, immunological, and virological effects of HHART on HIV-infected children in Mombasa, Kenya, data were taken and analyzed from 29 children. The result reveals that weight for age and CD4 cell have increased stepwise, while viral load decreased from a baseline by a factor more than 17 times (Song et al.2007).

Another study at primary care clinics in Lusaka, Zambia, held to demonstrate the clinical and immunological outcomes on 2938 children enrolled in a paediatric treatment program who started ART have verified that mortality rate was significantly associated with: CD4 cell depletion, lower weight for age, younger age, and anemia. And the mean CD4 cell percentage at ART initiation for more than 53% of children who had at least one repeat measurement has shown an increasing trend as measured every six months after initiation of ART (Bolton Moore, 2007).

The introduction of ART presents an enormous opportunity in terms of reducing morbidity and mortality due to AIDS worldwide. Ethiopia has been engaged in the scale up of ART access to its people since 2005. The free ART program was launched in July 2005. Despite the many challenges, ART scale up has recorded the greatest achievement over the last few years. The service has been expanded from only three health facilities in 2005 to 400 in 2008. The number of people who started on ART has also shown an unprecedented increase during the same period from 900 in 2005 to 180,447 by end of December 2008 (Seyoum et al. 2009).

A study by Selamawit (2009) on 423 patients identified factors on those patients at a risk of treatment failure has shown that the mean survival time (without treatment failure) was 53 months. Females were found to have a higher survival time of 57 months and males have significantly higher risk of developing treatment failure. Those with two or more episodes of poor adherence during their follow-up have a significantly higher failure compared to those with no episode of poor adherence. Missed appointment is another independent predictor of treatment failure. The study has shown that non-adherence to medication and clinic visits are associated with treatment failure. Following patients closely for their level of adherence and their trend of

missing clinic visits can be used to help identify those at higher risk of treatment failure. Providing intense adherence counseling for these patients may prevent occurrence of failure.

A research conducted by (Yonas, 2005) assessed the degree of adherence with antiretroviral therapy, identify which factors influence it, and described the everyday experience of PLWHA on ARV therapy. The researcher were used a combination of methodologies, including questionnaires, interviews and medical record review with patients in selected hospitals. Data on drug adherence were collected using patient self-report and depression was measured using Beck's depression inventory (BDI). Clinical data were recorded by asking the patient and reviewing their chart. Knowledge about ART was assessed by questions presented in "Yes" or "No" format.

Finally from a total of 431 HIV infected patients responded to the survey questionnaire. 81.2% of patients were adherent by self-report in the week before the assessment. The major reason reported for non-adherence were, being too busy with other things or by simply forgetting (33.9%) and being away from home (27.5%). Correlates of adherence in the multivariate analysis controlling for socio-demographic differences were: having regular follow-up, not being depressed, having no side effects, fitting a regimen to the daily routine, being satisfied with the relationship with health care providers, and the perception that doctors were capable and had access to assistance and reliable pharmacy.

Another study by Binyam (2008) that has been undertaken to assess the impact of malnutrition in survival of HIV infected children after initiation of ART based on 475 HIV infected children has

shown that CD4 count, hemoglobin value and weight for age were baseline predictors despite the obvious benefit of ART use on HIV related survival.

According to the studies above, the survival time of HIV infected patient after initiation of ART is a function of baseline variables like CD4 count, clinical stage of the disease, weight, age, drug adherence, types of treatment (mono-therapy, bi-therapy or triple therapy), viral load, nutrition, hemoglobin value and so on. This study is designed to 'identify additional' factors that affect the survival time of HIV infected children after initiation of antiretroviral treatment (ART) as a case study in Felege-Hiwot hospital ART unit at Bahir Dar city.

CHAPTER THREE

3. DATA MINING CONCEPTS AND TECHNIQUES

3.1. Data mining and knowledge discovery

It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% (Witten and Frank, 2000). As the volume of data increases, the proportion of information in which people could understand decreases substantially. This reveals that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which results in increasing information gap. Consequently, scholars begin to realize this bottleneck and to look into possible remedies. Current technological progress permits the storage and access of large amounts of data at virtually no cost. Although many times preached, the main problem in a current information-centric world remains to properly put the collected raw data to use (Kurgan and Musilek, 2006). The true value is not in storing the data, but rather in our ability to extract useful reports and to find interesting trends and correlations, through the use of statistical analysis and inference, to support decisions and policies made by scientists and businesses (Fayyad et al. 1996). To bridge the gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as Data Mining (DM)or Knowledge Discovery in Databases (KDD) has emerged in recent years.

Different scholars provided different definitions about DM. According to Berry and Linoff (2004); Han and Kamber (2006), DM is the process of extracting or “mining” knowledge from large amounts of data in order to discover meaningful patterns and rules. Witten and Frank

(2000) have also noted that DM is valuable to discover implicit, potentially useful information from huge data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns.

DM involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets (Two Crows Corporation, 1999). These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, DM consists of more than collecting and managing data; it also includes analysis and prediction and use of algorithms that improve their performance automatically through experience, such as neural networks or decision trees.

According to Han and Kamber (2006), the major reason that DM has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. DM tools perform data analysis and may uncover important data patterns. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration (Han and Kamber 2006).

DM is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high end computing, and others (Han and Kamber, 2006). DM methodology often can improve upon traditional statistical approaches for solving business solutions by finding additional, important variables, by identifying interaction among terms and detecting nonlinear relationships (SAS

Institute Inc. 1999). Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs.

3.2. Mining Algorithms

3.2.1. Decision Tree

A decision tree is a classifier expressed as a recursive partition of the instance space (Lior, 2008). The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called a "root" that has no incoming edges (Lior, 2008). All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an "internal" or "test" node (Lior, 2008). All other nodes are called "leaves" (also known as "terminal" or "decision" nodes). In the decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values (Lior, 2008). In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector (affinity vector) indicating the probability of the target attribute having a certain value (Lior, 2008).

3.2.1.1. J48

The basic algorithms for decision tree induction is a greedy algorithm which constructs decision trees in a top down approach dividing each node recursively until a leaf node is encountered (Ian, 2005). The following algorithm shows the generation of a decision tree from a raining tuples of data partition (Ian, 2005).

Input

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

Output

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

To construct optimal decision tree, Entropy and Information Gain needs to be calculated. The information gain measure enables to select the test attribute at each node in the tree and the attribute with the highest information gain or greatest entropy reduction is chosen as the test attribute for the current node (Jiawei et.al 2006).

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (C1, C2, C3 . . . , Cm). Let Si be the number of sample of S in class Ci. The expected information needed to classify a given sample is given by:

$$I(s1, s2, \dots, sm) = -\sum_{i=1}^{m} Pi \log_2(Pi) \quad \text{..... (3.1)}$$

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

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

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

Gain (A) = I (S_1, S_2, S_3, …., Sm) – Entropy (A) ……. (3.3)

This algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly (Jiawei et.al 2006). Decision tree use the above formulas to determine

which attribute to split on. The highest information gain considered as the test attribute for the given database (Jiawei et.al 2006). Then, a node is created and branches out. The same procedure followed for the next attribute to split.

3.2.2. Rule Induction

There is an alternative approach to rule induction that avoids global optimization but nevertheless produces accurate, compact rule sets (Ian, 2005). The method combines the divide and conquers strategy for decision tree learning with the separate-and-conquer one for rule learning. it adopts the separate-and-conquer strategy in that it builds a rule, removes the instances it over, and continues creating rules recursively for the remaining instances until none are left (Ian, 2005).

3.2.2.1. PART

PART (Partial Decision Tree) is a rule induction algorithm which grabs rule from a decision tree. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees (Ian, 2005). To generate such a tree, the construction and pruning operations are integrated in order to find a “stable” sub tree that can be simplified no further (Ian, 2005). Once this sub tree has been found, tree building ceases and a single rule is read off. The following algorithm depicts the steps and procedures followed in implementing PART rule induction.

Initialize E to the instance set

For each class C, from smallest to largest

BUILD:

Split E into Growing and Pruning sets in the ratio 2:1

Repeat until (a) there are no more uncovered examples of C; or (b) the description length (DL) of rule set and examples is 64 bits greater than the smallest DL found so far, or (c) the error rate exceeds 50%:

GROW phase: Grow a rule by greedily adding conditions until the rule is 100%accurate by testing every possible value of each attribute and selecting the condition with greatest information gain G

PRUNE phase: Prune conditions in last-to-first order. Continue as long as the worth W of the rule increases

OPTIMIZE:

GENERATE VARIANTS:

For each rule R for class C,

Split E afresh into Growing and Pruning sets

Remove all instances from the Pruning set that are covered by other rules for C

Use GROWS and PRUNE to generate and prune two competing rules from the newly split data:

R1 is a new rule, rebuilt from scratch;

R2 is generated by greedily adding antecedents to R.

Prune using the metric A (instead of W) on this reduced data

SELECT REPRESENTATIVE:

Replace R by whichever of R, R1 and R2 has the smallest DL.

MOP UP:

If there are residual uncovered instances of class C, return to the BUILD stage to generate more rules based on these

CLEAN UP:

Calculate DL for the whole rule set and for the rule set with each rule in turn omitted; delete any rule that increases the DL Remove instances covered by the rules just generated.

Continue…..

3.3. Confusion Matrix

Confusion matrix is a tool for analyzing how well the classifier can recognize tuples of different classes (Ian, 2005). Give m classes, a confusion matrix is a table of at least m by m. For instance, in two class case with classes yes and no, sick or healthy, cancer or not cancer, live or dead, or lend or not lend, a single prediction has four different possible outcomes.

Table 3.1: Different outcomes of a two class prediction

		Predicted Class	
		Yes	No
(Actual Class)	Yes	TP	FN
	No	FP	TN

In the above table 3.1, four possible outcomes of a two class classifier are observed, where true positive (TP) and true negative (TN) are the correct classifications. A false positive (FP) is when the outcome is incorrectly predicted as yes (or positive) when it is actually no negative). A false negative (FN) is when the outcome is incorrectly predicted as negative when it is actually positive (Ian, 2005), (Jiawei, 2006). Confusion matrix enables the mechanism to understand how well the classifier has classified the tuples as "yes" and "no". On the other hand sensitivity (the true positive rate) and false positive rate can also be computed using this confusion matrix to rate the performance of a classifier (Jiawei, 2006). The equation 3.4 and 3.5 placed below has been used to compute the sensitivity and false positive rate of a classifier respectively.

$$\text{TPR} = \frac{TP}{TP+FN} \quad \text{.....(3.4)}$$

$$\text{FPR} = \frac{FP}{TN+FP} \quad \text{.....(3.5)}$$

The other parameter used in this study to compare the performance of classifiers is “precision”; which indicates the percentage of instances classified as positives by the learned model and that are actually positives. The equation (3.6) presented below can be used to compute the precision of a given classifier.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{.....(3.6)}$$

The F-value metric is one measure that combines the trade-offs of precision and recall, and outputs a single number reflecting the "goodness" of a classifier in the presence of rare classes (Nitesh, 2012). The F-value represents the trade-off among different values of TP, FP, and FN. The expression for the F-value is as follows:

$$\text{F} - \text{value} = \frac{((1+\beta)^2 * recall * precision)}{(\beta)^2 * recall * precision} \quad \text{..... (3.7)}$$

Where corresponds to the relative importance of precision versus recall. It is usually set to 1. The entire success of the classifier or accuracy of the classifier is the number of correctly classified instances divided by the total number of instances. On the other hand the error can be computed through subtracting the accuracy from one (Ian, 2005). The overall accuracy of the classifier can be computed by using the equation 3.8 below.

$$\text{Over all accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+FN}} \quad \text{.....(3.8)}$$

In this research the performance of each model is weighted in terms of sensitivity, FPR, F-measure, and accuracy values. Based on the remark from this comparison, the one with better performance was discussed in brief.

3.4. ROC Curve

ROC curves are a useful visual tool for comparing two classification models. The acronym stands for receiver operating characteristic, a term used in signal detection to characterize the tradeoff between hit rate and false-alarm rate over a noisy channel (Jiawei, 2006). ROC curves depict the performance of a classifier without regard to class distribution or error costs (Jiawei, 2006). They plot the true positive rate on the vertical axis against the false positive rate on the horizontal axis (Jiawei, 2006). The former is the number of positives included in the sample, expressed as a percentage of the total number of positives (TP Rate = 100 × TP/ (TP + FN)); the latter is the number of false positives included in the sample, expressed as a percentage of the total number of negatives (FP Rate = 100 × FP/ (FP + TN)). A sample ROC curve representing the percentage of true positives and false positives is presented in the figure 3.1 below. The plot also shows a diagonal line where for every true positive of such a model, there is more likely to encounter a false positive. Thus, the closer the ROC curve of a model is to the diagonal line, the less accurate the model. If the model is really good, initially we are more likely to encounter true positives as we move down the ranked list. Thus, the curve would move steeply up from zero. Later, as we start to encounter fewer and fewer true positives, and more and more false positives, the curve cases off and becomes more horizontal.

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

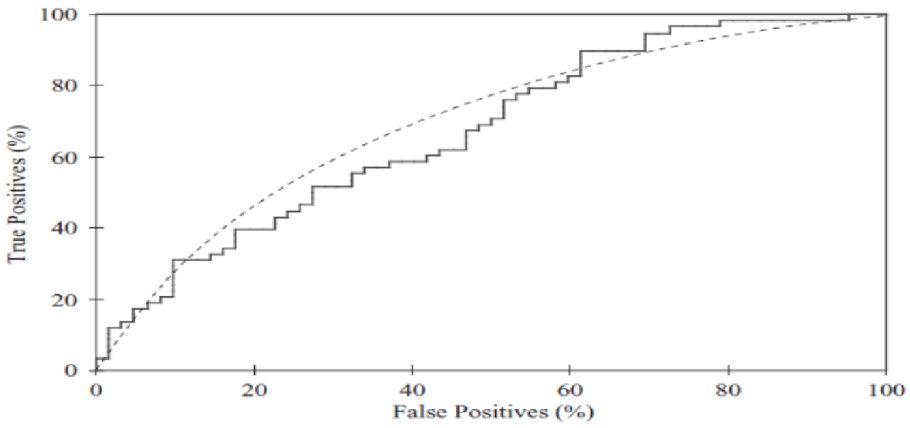

Figure 3.1: Sample ROC Curve

3.5. The Data Mining Process

DM requires massive collection of data to generate valuable information (Deshpande and Thakare 2010). The data can range from simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. Deshpande and Thakare indicated that the data retrieval is simply not enough to take complete advantage of data. It requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

A typical DM process includes data acquisition, data integration, data exploration, model building, and model validation (Deshpande and Thakare, 2010). Both expert opinion and DM techniques play an important role at each step of this knowledge discovery process.

3.5.1. Data acquisition

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

3.5.2. Data pre-processing

Once the target data is selected, the data is then pre-processed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this pre-processing step, researchers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes. Then data is transformed to reduce the number of variables by converting one type of data to another (e.g., numeric ones into categorical) or deriving new attributes.

3.5.3. Building model

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

technique with the overall objective of DM. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

3.5.4. Interpretation and model evaluation

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

3.6. Data Mining Function

Data mining is utilized for the intention of finding of hidden information in a database upon developing of model which could best fit the data. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. The ability to extract useful knowledge hidden in the data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer based methodology including new techniques for discovering knowledge from data is core function of data mining. It searches for new, valuable, and nontrivial information in large volumes of data (Mehamed, 2003). Data mining tasks are in general classified in to two main categories (Mehamed, 2003): predictive-oriented and descriptive oriented. Predictive data mining tasks produce the model of the system described by the given dataset to build a model that permits the value of unknown variable to be

predicted from the known values of other variables (Ruben, 2009). It is a technique that involves using some variables or fields in the dataset to predict unknown or previously unseen future values of other variables of interest. It is usually used to create a model based on a set of predictors to relate the dependent variables. Examples of predictive modeling includes classification, prediction etc.

The second category of data mining function is descriptive mining task. This is another data mining task used to characterize the general properties of the data in the database (Han and Kamber, 2006). It produces new, nontrivial information based on the available dataset and is to gain an understanding of the analyzed system by uncovering patterns and relationships in large datasets. The goal of a descriptive model is to describe all of the data or the process generating the data (Ruben, 2009). Examples for descriptive data mining are clustering, summarization, association rule discovery, and sequence discovery. The followings are some of the examples from both data mining tasks how they are working in real pattern discovery process.

3.6.1. Classification

Classification is one of the predictive data mining tasks. It is a technique used to predict group membership for data instances by assigning previously unseen records a class as accurately as possible. It is said to be the process of finding a model or function that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown (Han and Kamber,2006).

The derived model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision

trees, mathematical formulae, semantic network etc (Han and Kamber,2006). Each technique employs a learning algorithm to identify a model that best fits the relationship between the attributes set and the class level of the input data.

After having an accepted accuracy level, one can use the model for classification of new data tuples. Applications examples of classification in health sectors are the following (Han and Kamber, 2006).

- A hospital may want to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness.
- Classifying the type of drug a patient should be prescribed based on certain patient characteristics in hospital.
- A medical researcher wants to analyze breast cancer data in order to predict which one of the specific treatments a patient should receive.

There are various classification algorithms; among which the main ones are the following (Han and Kamber, 2006).

3.6.1.1. Decision tree induction

When decision tree induction is used for attribute subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples. Thus, the algorithm choose the best attribute to partition the data into individual classes includes ID3, C4.5, and CART (Mehamed, 2003).

In decision tree construction, selection of splitting attributes is necessary in order to void irrelevant attributes by examining the effect of each attribute for the distinct class and its likelihood for improving the overall decision performance of the tree, since the feature with minimum impact on dependent variable may distort the trees performance and the classification accuracy. There should be certain requirements before decision tree algorithms become applied (Mehamed, 2003).

First: since decision tree algorithms represent supervised learning, they require pre-defined target variables and training dataset which provides the algorithm with the values of the target variable.

Second: this training dataset should be rich and varied, providing the algorithm with a healthy cross-section of the types of records for which classification may be needed in the future. Decision trees learn by example, and if examples are systematically lacking for a definable subset of records, classification and prediction for this subset will be problematic or impossible.

Third: the target attribute classes must be discrete i.e. one cannot apply decision tree analysis to a continuous target variable. The target variable needs to take on values that are clearly demarcated as either belonging or not belonging to a particular class. One of the most attractive aspects of decision trees lies in their interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf (Mehamed, 2003). Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule (Mehamed, 2003).

The challenge with decision tree is over fitting. As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex (Han and Kamber,

2006). This potentially leads to the concept of overfitting which consequently brings the notion of pruning; this implies removing of branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree (Han and Kamber, 2006).

3.6.1.2. Rule based classification

Though the decision tree is a widely used technique for classification purposes, another popular alternative to decision trees is classification rules which can be expressed as paths IF-THEN rules so that humans can understand them easily (Getu, 2007). A rule-based classifier uses a set of IF-THEN rules for classification; it is a relationship between antecedent, and consequent i.e. an expression of the form IF condition THEN the conclusion.

The algorithm decision tree is the best known method for deriving rules from classification trees (Bramer, 2007). For example, one could have the following set of rules to classify the weather condition. If temperature < 50°F, then weather = cold. If temperature > 50°F AND temperature < 80°F, then weather = warm. If temperature > 80°F, then weather = hot (Berry , 2006). Although any of the logical expressions are allowed, preconditions are usually connected with the AND operation. The advantage of IF-THEN rule is the rules are order independent i.e. regardless of the order of rules executed, the same classification of the classes is possible to reach (Berry, 2006). The challenges is the generated rules are often more complex than necessary and contain redundant information and the rules generated this way may be unnecessarily complex and incomprehensible (Berry, 2006).

3.7. Methodology of Data Mining Research

One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets (Mehamed, 2003). Data mining tools perform data analysis and uncover important data patterns, contributing greatly to different business strategies including medical researchers. The widening gap between data and information calls for a systematic development of data mining tools that will turn data tombs into golden nuggets of knowledge. Thus, patterns and knowledge from data mining is using for sound judgment and proactive decision making in different organization including health care sectors.

Broadly used methodologies in data mining are KDD (Knowledge Discovery in Data base), CRISP-DM (Cross-Industry Standard Process for Data Mining), SEMMA (Sample Explore Modify Model Assess), and HYBRID process (Ana, 2008).

3.7.1. Knowledge Discovery in Database (KDD)

The first KDD process was proposed by Fayyad in 1996(Ana, 2008). This process consists of several steps that can be executed iteratively. KDD has been more formally defined as it is non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is the process of knowledge discovery while data mining is a technique applied or knowledge discovery considered as just a step in the entire process (Ana, 2008). As shown in Figure 3.2, the KDD process consists of five steps: data selection, data preprocessing, data transformation, data mining and interpretation/evaluation

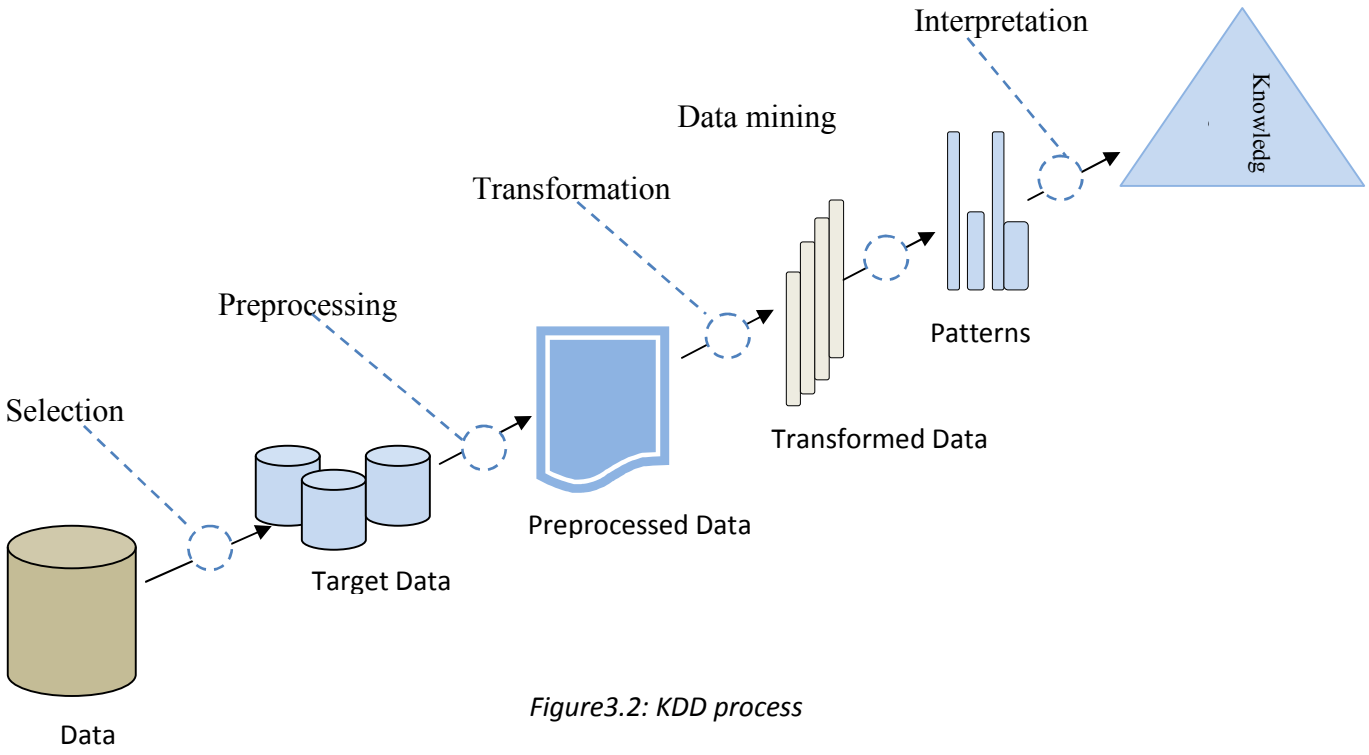

Figure3.2: KDD process

Given data, the first step in KDD is data selection. In this stage creating a target dataset on focus of a subset of variables needed on which discovery aimed to solve the problem are selected. For discovery purposes, data relevant to the analysis task are retrieved from the database and unnecessary data attributes should be removed.

In order to produce effective data mining models in terms of quality and performance, the raw data need to undergo preprocessing in the form of data cleaning. Because real world data are mostly dirty and unclean which need to correct bad data that encountered from data redundancy, incompleteness or missing attributes value, noise, and inconsistency in order to make knowledge searching paths ease for mining algorithms. Therefore, data quality needs to be assured in this step before ahead to next phase of knowledge discovery process in data mining. Because of the

use of different sources, data that is fine on its own may become problematic when we want to integrate it. In this step data need to be combined from multiple sources, such as database, data warehouse, files and non-electronic sources into a coherent store. We need to merge different sourced data by keeping uniform format for all before running data mining tools and techniques. During transformation phase, data are consolidated into forms appropriate for mining to reduce data size by dividing the range of data attribute into intervals each containing approximately same number of samples or to scale attribute data to fall within a specified range. Therefore, values of attributes are changed to a new set of replacement values to ease data mining. Data mining is the next essential process where intelligent methods are applied in order to extract hidden patterns in the data. This phase requires analysis of the main problem for patterns of interest in the data depending on the business objectives and data mining requirements. Different data mining algorithms and techniques are used for searching knowledge or interesting patterns to construct predictive or descriptive models.

Model creation is followed by performance evaluation which measures the accuracy rate of the system. The mined pattern enables to identify the truly interesting ones. For any errors or mismatched result generation as compared to domain area perspectives, the process restarts to initial step so as to provide accurate results. Accuracy means the percentage of test set samples that are correctly classified by the classifier. Finally, visualization and knowledge representation are used to present the mined knowledge to the users and stored as new knowledge in the knowledge base. Incorporating the knowledge in to another system for implementation purpose, documentation and report for presenting the benefit of the knowledge to interested parties,

incorporating the knowledge with previously known knowledge in the area are some of the important activities during this phase.

3.7.2. CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP–DM was developed in 1996 by analysts for fitting data mining into the general problem solving strategy of a business or research unit (Milley, 2000). CRISP–DM is one of the most widely used methodologies in extraction of knowledge which has a life cycle consisting of six phases which is an iterative and adaptive process (Cios, 2007), as depicted in Figure 3.3

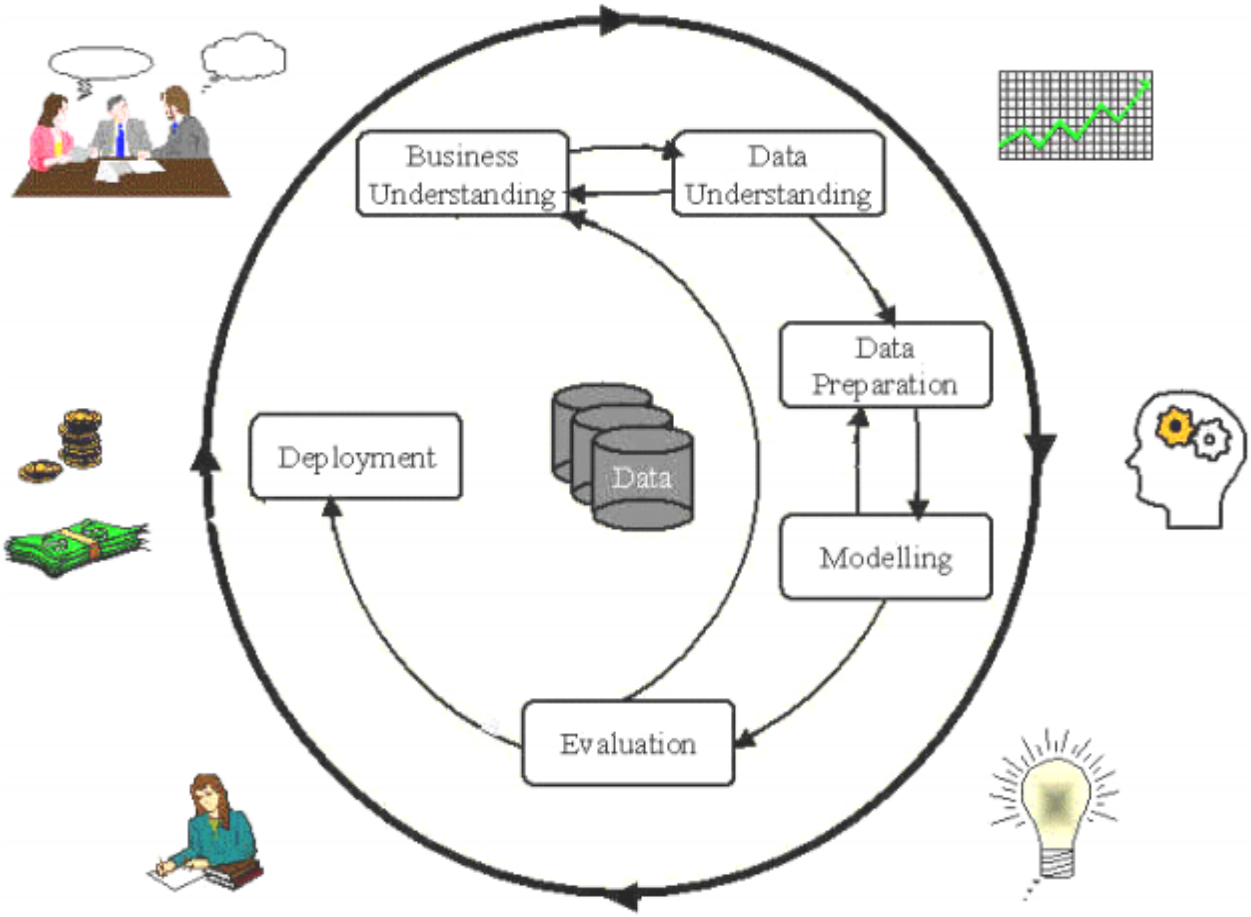

Figure3.3: CRISP-DM knowledge discovery Process Model

In CRISP- DM, the sequences of the phases are adaptive i.e. the next phase in the sequence often depends on the outcomes associated with the preceding phase. Business understanding is the initial phase in the CRISP–DM standard process which focuses on understanding business area in which DM objectives and project requirements are assessed as a whole from business perspective points of view. It also translates these goals and objectives into the formulation of a data mining problem definition and prepares a preliminary strategy for achieving the desired objectives. Further, it is broken in to determination of what clients really wants to accomplish from business perspectives, assessment of the situation for fact finding about the resources, constraints and assumption, determination of data mining goals, and states project objectives in technical term and finally description of the project plan for achieving the data mining and business goals. Once the business is well defined and understood, data understanding phase begins with collecting the initial data and continues with several activities in order to become familiar with the data that helps to identify quality of the data. During this step, the following tasks are performed. First, collecting initial data for modeling; second, data description to get insights through descriptive statistics available in the statistical tools; third, exploration of data to capture an overall sense of the dataset through computing summary. Lastly, verification, and visualization of data quality is checked if any unnecessary data fields with incomplete, inconsistent, noisy, and redundant values existed.

The data preparation step contains all activities needed to construct the final dataset. It starts from preparing the initial raw data to the final dataset which is ready for application of data mining tools. This step further divided in to four steps. The first step is data selection that is appropriate for analysis. This is followed by data cleaning which is making data ready for the

modeling tools. Data construction is the third step that attempts to produce derived attribute, new records and transformed values for existing attributes. Then, data integration is to combine data from multiple sources (records or tables), and, finally data formatting for reconstructing data values without changing its meaning.

After data being ready to apply data mining tools in proceeding step, various modeling techniques are selected and applied at this phase. Since some data mining tools may require specific formatting for input, it may needs reiteration into the previous phases for improvement. The modeling step selects first modeling techniques based on data mining objectives and also generates test set to evaluate and validate model performance. This is followed by model building using the modeling tool. Finally, assess and interpret the pattern according to the domain knowledge. After models have been built, they need to be evaluated to check whether they fulfill the requirements and objectives set at the beginning of the project. The model is also evaluated from the point of business objectives. Reviewing of steps executed to build the model and evaluating the models for quality and effectiveness before generating them for end users in the field are also performed. At the end, decision regarding the deployment and use of the data mining results is reached.

3.7.3. SEMMA

Another well-known methodology developed by the SAS institute is SEMMA (Sample, Explore, Modify, Model, and Assess) which refers to the process of conducting a DM project as depicted in Figure 2.3.

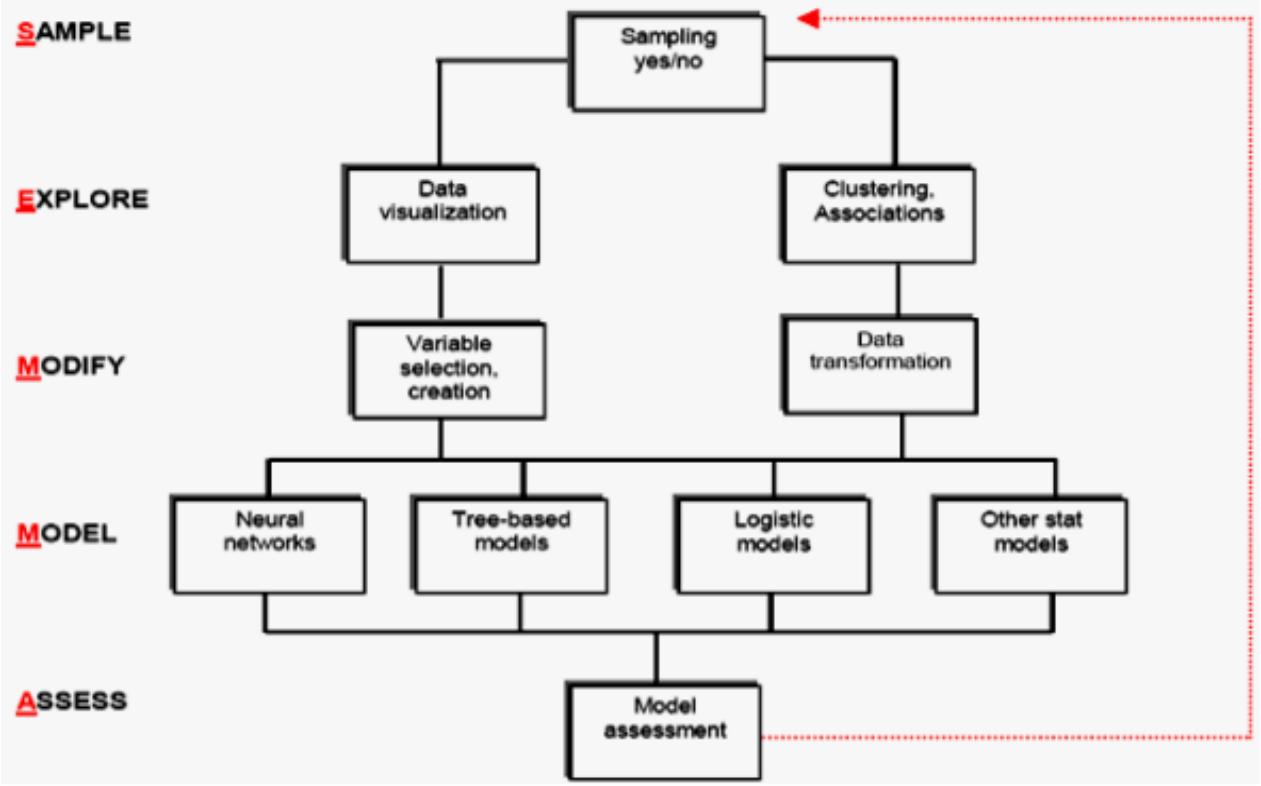

Figure 3.4: SEMMA Process model

The first phase in SEMMA process model is sampling. In this phase, a portion of a large dataset is extracted in order to take reliable and statistically representative sample from the huge data for optimal cost and computational performance. Sample data selection is followed by explore. This is a state where searching for unanticipated trends and anomalies in order to gain a better understanding of the dataset occurs. This helps to refine the dataset and redirect the discovery process. During the modify phase, user creates, selects, and transforms the variables upon which to focus the model construction process. In addition, it manipulates data to include information and handle outliers to increase significance of variables and focus on model selection process. Once the data is prepared, the modeling phase constructs models that explain patterns in the data by applying modeling techniques in data mining. The modeling techniques are selected based on

the objectives of the data mining project. This is the assess phase in which the usefulness and reliability of the model is evaluated from the data mining process and how well it performs. A common means of assessing a model is to apply it to a portion of dataset put aside for testing during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, one can test the model against known data. By assessing the outcome of each stage in the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. In SEMMA, the sample steps goes equivalently with selection step of KDD and continues till to last assessment phase as interpretation/evaluation of the discovered knowledge in KDD. However, KDD manifests the pre-KDD and Post KDD that SEMMA does not.

3.7.4. Hybrid model

The development of academic models such as the nine-step model and eight-step model and industrial models such as five-step model and the six-step CRISP-DM model has led to the development of hybrid model that combines aspects usable for DM research. It was developed by (Cios et al, 2007) based on the CRISP-DM model.

Hybrid process is characterized by providing more general, research oriented description of the steps. The hybrid model also encourages the application of knowledge discovered for a particular domain in other domains and it has a six step process as depicted in Figure 3.5 (Cios et al, 2007).

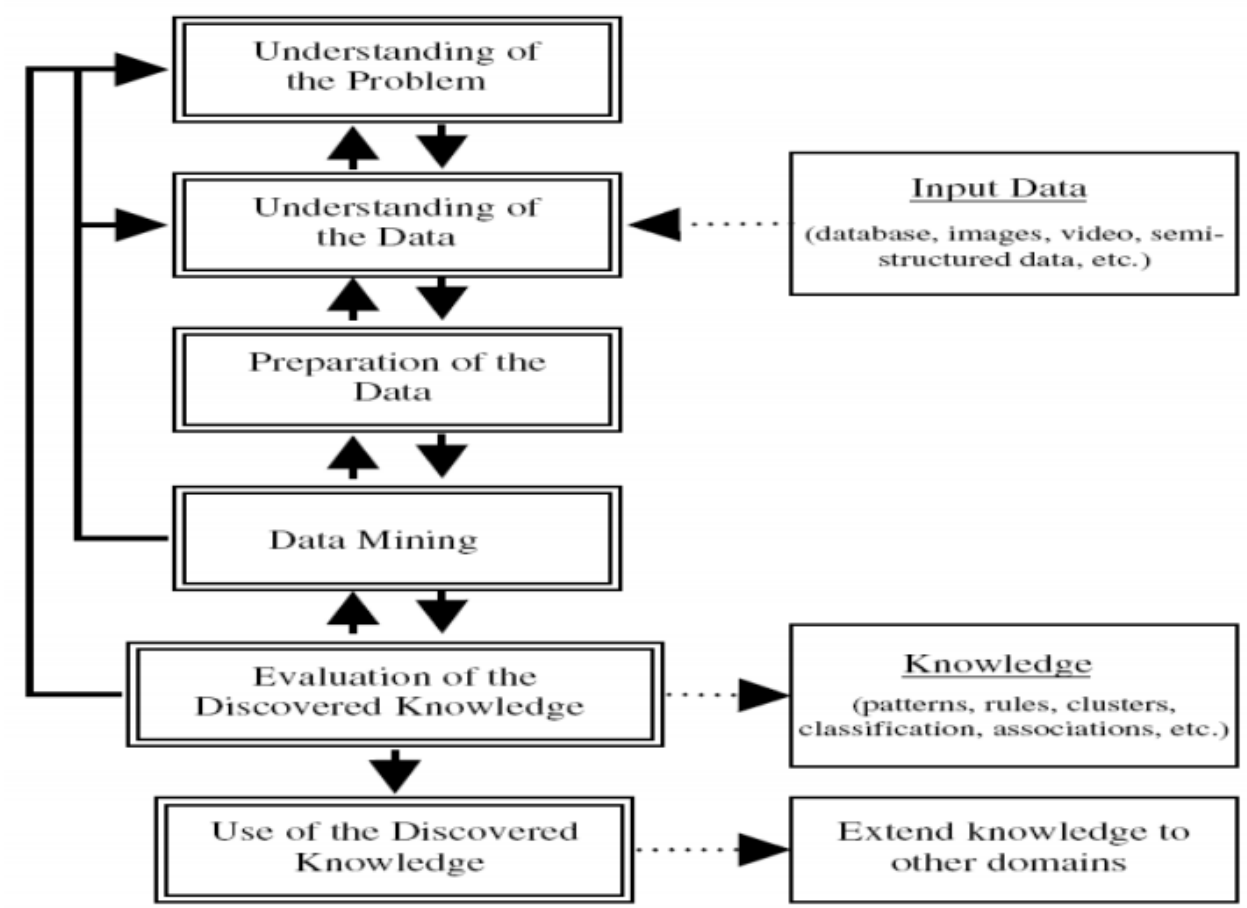

Figure 3.5: Hybrid-DM Process Model

The Hybrid-DM model presented in Figure 3.5 above consists of a six-step knowledge discovery process which includes; understanding the problem domain, understanding data, preparation of data, data mining, evaluation of the discovered knowledge, use of the discovered knowledge (Cios et al, 2007).

Summary of correspondences between KDD, SEMMA, CRISP-DM, AND HYBRID models are presented in Table 2.1(Ana, 2008).

Table 3.2 Summary of data mining models

KDD	***SEMMA***	***CRISP-DM***	***HYBRID***
Pre KDD	----------	Business understanding	Problem domain understanding
Selection	Sample	Data understanding	Data understanding
Preprocessing	Explore		
Transformation	Modify	Data preparation	Data Preparation
Data mining	Model	Modeling	Data mining
Interpretation/evaluation of the discovered knowledge	Assessment	Evaluation	Evaluation
Post KDD	----------	Deployment of discovered knowledge	Use of discovered knowledge

From the Table 3.2 by doing a comparison of the models, some of them follow same steps to discovery process while others follow different steps. For example in KDD and SEMMA stages the first approach is equivalent. Sample can be identified with Selection; Explore can be identified with Preprocessing; Modify can be identified with Transformation; Model can be identified with Data Mining; Assess can be identified with Interpretation/Evaluation.

3.8. Related Work

There are a number of researches done to apply data mining techniques in health care domain in general and HIV/AIDS intervention programs in particular. A study done by Elias indicated that data mining techniques can be used to develop HIV status predictive model. In that study, J48 and ID3 decision tree implementation classification algorithms were used. And also Apriori association rule generation algorithm was used to show the existence of association between the dependent variable HIV status and the independent variables. The study has used 51,270 records and 17 selected attributes to train and test the models. Moreover, the study is organized in to five experimental scenarios for both J48 and ID3 classifiers to build up the respective models. In

addition to the classification mining tasks, association rule discovery was made to uncover the underlying association between the predicted variable HIV status and the predicting variables. The findings of Elias study have shown that, the classification rules revealed that females are more vulnerable to HIV than Males. The other two classification rules were generated in terms of age group of clients. According to the study, age group 25-49 were the most susceptible subset of the population and age group 50 and above were also becoming vulnerable to HIV/AIDS as the patterns have indicated. Group of the association rules indicated that there is a direct relationship between never married clients and HIV negative status. The other observed association rule is those clients whose primary reason to visit the center labeled risk is associated with positive HIV status. Moreover the study revealed that the attributes age, economic status, and knowledge about the disease are among the contributing factors to engage in risk which may amount to HIV positive result.

Teklu (2010) has attempted to investigate the application of data mining techniques on antiretroviral treatment (ART) service with the purpose of identifying the determinant factors affecting the termination /continuation of the service. This study applied classification and association rules using, J48 and apriori algorithms respectively. The study was conducted using 18,740 ART patients' dataset. The methodology employed to perform the research work is CRISP_DM. Finally, the investigator proved that the applicability of data mining on ART by identifying those factors causing the continuation of the service.

Birru (2009) has also investigated the applicability of data mining on VCT taking the case of CDC. He used 56,486 dataset from 2002 to 2008. The dataset contains unbalance HIV positive and negative clients' data and after the dataset is balanced only 14,793 records are considered for

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

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

Shegaw (2002) has investigated how social, economic, behavioral, and environmental and health related factors affect child mortality at the district of butajira by applying data mining technology. This study was used neural network and decision tree algorithms. The study was conducted using 1,100 records of children were used to build and test both neural network and decision tree models. The methodology employed to perform the research work is KDD. Finally he proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality.

Another study has been conducted by Amanuel (2009) using data mining techniques to predict household health seeking patterns using BRHP dataset. The researcher aim was to develop a model that identifies risk factors and patterns of household health seeking behavior at Butajira district. He used a total of 60,446 records for experiments with implementation of J48 decision tree techniques. The finding of the researcher indicated that with an accurate rate of 89.9017%, predicting household health seeking pattern through data mining techniques is possible.

A research conducted by Misganaw (2013) on mining ART data set to predict CD4 cells count applying data mining techniques to build CD4 count predictive model using socio-demographic, clinical and biological features. The study was mainly focused on developing a model capable of predicting CD4 cell count of patients after six, twelve, and eighteen months of treatment.

The activities have been guided by a Hybrid-DM model which is a six step process model. The study has used 7,252 instances, ten predicting and three outcome variables to run the experiments. The study has selected classification mining to build the classifier model and the

mining algorithms were j48, PART, SMO and MLP used to run different experiments. Ten-fold cross validation technique has been used to train and test the classifier models. Performance of the models were compared using accuracy, TPR, FPR, F-measure, and the area under the ROC curve.

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

In all the researches done, different researchers tried to develop a model and searched new knowledge that may help objectives of Anti retroviral therapy (ART) using ART dataset as information source.

Previously there were research works that have been carried in application of data mining techniques using ART dataset as an important information source. But, to the knowledge of the researcher, no previous researches have been done to predict survival rate of HIV/AIDS patients by applying data mining techniques in the area. Thus, this research has a great contribution to generate patterns that help in planning a better strategy and effective decision making for HIV/AIDS programs.

CHAPTER FOUR

4. DATA PREPARATION

Data Mining is a technology that uses various techniques to discover hidden knowledge from a data stored in large databases, data warehouses and other massive information repositories (Jiawei and Micheline, 2006). To discover non-trivial knowledge and patterns, the database must undergo effective data preparation to bring a valid output.

The data mining model used in this study was a Hybrid-DM, which is a six step knowledge discovery process. Among these phases, the first three, understanding problem, data understanding and data preprocessing are meant to prepare the data for data mining tasks. As noted in Jiawei and Micheline (2006), data understanding and preprocessing usually consumes the majority of the effort in the entire data mining process.

In any data mining task the first step is to get clear understanding of the problem to be solved. In this chapter data understanding activities such as; data collection, data description and data formatting have been undertaken. Secondly, the data has been pre -processed by employing data cleaning and data selection techniques.

4.1.Business Understanding

Business understanding is a stage where the researcher is acquainted with the overall business process or working conditions of the identified work process. To this end, the researcher has made efforts to understand what the business is and its' in and out as much as possible. This study was done in consultation with domain area experts to have an in-depth insight into the

problem domain and also physical observation of the situation was done to have a real picture of the problem. The domain area experts constitute physicians, nurses who are engaged in giving counseling services and taking care to those admitted to inpatient department and ART data clerks.

4.2.Data Understanding

Basically data is crucial element in data mining research. Without well understanding of data, data mining is unthinkable. Therefore, for the success of a data mining research data understanding is a mandatory step to bring about a valid result. In order to meet the general objective of this research, collecting of ART (antiretroviral therapy) data is a prerequisite.

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data in order to identify data quality problems, discover first insights into the data, or detect interesting subsets to form hypothesis for hidden information (The CRISP-DM consortium, 2000).

Data understanding begins with collection of the initial dataset. The data collection was done in Jimma University hospital which is located in the south western part of Ethiopia. Somewhat helpful discussions were made with domain area experts and physicians working on ART clinic within the mentioned hospital in order to select the final dataset to be used for the study. In addition to domain area experts, literatures done on related areas were analyzed to verify the selected attributes. Accordingly twelve (12) independent variables i.e. Age, Sex, Marital Status, Educational Status, Functional status, Eligible Reason, Religion, Occupation, Baseline WHO

stage, Drug regimen, Baseline CD4 count and one (1) outcome variable i.e. Status of patient were selected from the huge dataset.

4.2.1. Data Collection

As indicated in section 4.2, it is a bare fact that the concept of data mining doesn't exist without data. Collecting representative subset of ART data is a prerequisite to address the objective of this research. Therefore, Data were collected from the records of HIV patients who were in follow up treatment in the ART clinic of Jimma university specialized hospital ART Data bases. The Data base is in SQL server through which the data clerks can enter patient data and generate different reports. A full backup of the database of ART was taken from Jimma University specialized Hospital with raw data of 9200 records and 81 attributes. It stores data on HIV/AIDS patients' who are in care to start drug, Eligible to start date and on those who have already started using ART drug. The dataset stores eleven years record from the year 1996 to 2006. The total dataset obtained on ART is around 9200. As part of the curse of dimensionality, there were lots of missing values under some of the columns and there were also data items such as different IDs, Data values in date/time format which don't serve the purpose of mining. Moreover, there have been attributes which are not clear for what they stand for and which the expert who gave the data couldn't explain what they are for. Moreover, there were attributes containing redundant values such as birth date, birth date in Ethiopian Calendar, Age in Months, Age in Years, etc.

Therefore describing each of them at this point might not be necessary as there were many of them excluded as part of the tedious process the researcher followed to filter out the more relevant attributes that are important for analysis with the help of domain experts in the area and literatures done on related areas.

4.2.2. Attribute Selection

Records are evaluated and classified based on the values of their attributes. Of course, some of the attributes of a record may be irrelevant to the process of classification and thus should be excluded. Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean (Whitten and Frank, 2005).

Out of the 81 attributes of the original data set, 13 attributes (including the class attribute) which are believed by the domain experts to have significant contribution in predicting the survival rate of HIV/AIDS patients, which is the focus of this research, have been selected.

4.2.3. Instances Selection

Out of the 9,200 follow up patients, 7,951 cases were registered for ON ART follow up. Even from this number building a predictive model requires to give the learner algorithm with a training set that have all instance whose outcome or dependent attribute (class label) is not missing. Instance with missing values for outcome class are not useful for predictive model building in data mining because classification algorithms of data mining learn how instance were classified under the different classes. The classes are not existing means the algorithm learns nothing from these instance. As stated by Han and Kamber (2006), records without class labels (missing or not entered) should be ignored, provided that the data mining task involves classification. As this study uses classification algorithms for the purpose of predictive model building, the 251 records without class information were removed from subsequent analyses. The

remaining dataset were then having 7700 records whose outcomes are distributed in one of the two outcome categories.

4.2.4. Data Description

After selecting the relevant attribute, the next step is describing the data set. The ART dataset has eighty one attributes of text, number, and date & time formats. Out of 81 attributes 13 are selected based on their relevance for the research objective by consulting domain experts. The next table 4.1 shows the complete description of the whole attributes.

Table 4.1: Data description

No	Attributes	Description	Values	type
1.	Sex	The sex of patient	Male, Female	Nominal
2.	Age	The age of the patient in year	Numeric age values	Numeric
3.	Marital Status	Marital status of the patient	Never married, Married, Separated, Divorced, Widow	Nominal
4.	Educational Level	Educational level of the patient	No education, Primary, Secondary, Tertiary	Nominal
5.	Occupation	Patient's line of work	Employed, Unemployed, self employed	Nominal
6.	Religion	The religion of the patient	Muslim, Orthodox, Protestant, catholic	Nominal
7.	Eligible Reason	The reason why the patient attends ART	Clinical, CD4, TLC, WHO stage and the combination	Nominal
8.	Baseline CD4	The patients CD4 count	Numeric value	Numeric
9.	Functional Status	The patient's performance of daily activities.	W-working, A-Ambulatory, B-Bedridden, D-Delay	Nominal
10.	Baseline Weight	The patient's body weight	Number in Kg	Numeric
11.	Baseline WHO Stage	World Health Organization stage category of the patient	One, Two, Three, Four	Nominal
12.	Drug Regimen	The FIRST regimen the patient taken	1-a, 1-b, 1-c, 1-d, 1-e, 1-f, 1-g , 2-a, 2-b, 2-c, 2-d	Nominal
13.	Status	The life status of the patient either Alive or Dead	Alive, Dead	Nominal

4.2.5. Formatting the Data

Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last attribute being the outcome field the model to predict. It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute (The CRISPDM consortium, 2000).

The WEKA data mining tool requires the dataset to be in a comma separated file format called the Attribute Relation File Format (ARFF). The ARFF file format is the standard way of representing datasets that consist of independent, unordered instances and does not involve relationships between instances (Whitten and Frank, 2005). Hence, the ART dataset which was originally in a Microsoft SQL file format is extracted to Excel file format and then converted to an ARFF file format.

4.2.6. Data Preprocessing

4.2.6.1. Data Summarization

Descriptive data summarization techniques can be used to identify the typical properties of your data and highlight which data values should be treated as noise or outliers (Jiawei and Micheline, 2006). Furthermore, missing values can easily be identified through this technique which in turn facilitates the next phases of data preparation. To understand the nature of the data values in the selected ART dataset an exploratory data analysis was done for numeric. Descriptive statistics comprising of the total number of valid and missing instances, minimum, maximum, mean and standard deviation were computed to show the distribution of values for the numeric attributes.

In this study, three (3) numeric attributes were used. The following table 4.2 presents numeric variables and their distribution in the dataset.

Table 4.2: Exploratory Data Analysis

Exploratory Data Analysis				
		Age	Weight	Baseline CD4
N	Valid	7700	7700	7700
	Missing	0	0	0
Mean		29.66	41.83	131.23
Median		30.00	45	138.00
Mode		30	41	138
Std. Deviation		11.422	16.15	82.013
Minimum		0	1	1
Maximum		114	93	595
Percentiles	25 (Q1)	25	41	95
	50 (Q2)	30	45	138
	75 (Q3)	36	52	140
Outliers		< 8.5 and > 52.5	< 4.5 and > 83.5	>367.5

4.2.7. Data Cleaning

Data cleaning is a time-consuming and labor-intensive procedure in data preparation but one that is absolutely necessary for successful data mining (Witten and Frank, 2005). Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006). Thus, data cleaning has become a must in order to improve the quality of data so as to improve the accuracy and efficiency of the data mining techniques.

The data cleaning tasks performed to raise the data quality to the level required by the selected analysis techniques involves selection of clean subsets of the data and the insertion of suitable defaults. To this end, thorough discussion has been made with the domain experts and it is found out that missing attribute values at the time of data entry are recorded as “unknown” and for

those records the attribute is irrelevant they simply left it as a blank assuming that it would be obvious. To fix these problems some records with missing or unknown values for significant number of attributes are removed from the dataset. Noisy values for attributes are also deleted and set to blank.

4.2.7.1. Handling Missing Values

Missing data is a problem that continues to affect data analysis methods (Larose, 2005). The absence of information is rarely beneficial. All things being equal, more data is almost always better.

According to Larose (2005) missing values may occur for several reasons, such as malfunctioning measurement equipment, lack of consistency with other recorded data and thus deleted, or respondents in a survey may refuse to answer certain questions such as weight, height or religion and data may not be recorded due to misunderstanding. But those missing values needs to be given significant attention.

To deal with missing values, alternatives are suggested by Larose (2005) and Chackrabarti et al. (2009). These are:

- Ignore the missing value
- Replace the missing value manually
- Replace the missing value with a global constant to fill in the missing value
- Replace the missing value with some constant, specified by the analyst
- Replace the missing value with the field mean(for numerical variables) or the mode (for categorical variables)

In the study dataset, missing values were observed in both the numeric and nominal variables. Fortunately, out of nine nominal attributes only four (4) nominal attributes (which are Functional Status, Religion, WHO Stage, and drug regimen) has missing value. Hence, the percentage of missing values in the mentioned attributes is not considerably significant, thus the missing values were replaced by the mode for nominal attributes using Microsoft Excel version 2007 before applying data mining algorithms. The following table 4.3 depicts attributes with the number of missing values, mean, mode and the action taken to replace the missing values.

Table 4.3: Handling missing values

No	Attribute	Valid	Missing	Mean	Mode	Action taken
1.	Sex	7700	0		Female	No action taken
2.	Occupation	7700	0		unemployed	No action taken
3.	Eligibility reason	7700	0		CD4	No action taken
4.	Marital Status	7700	0		Married	No action taken
5.	Education Level	7700	0		Primary	No action taken
6.	Religion	7315	385 (5%)		Orthodox	Replaced by mode
7.	Functional status	7238	462 (6%)		working	Replaced by mode
8.	Baseline WHO stage	7238	462 (6%)		WHO stage 3	Replaced by mode
9.	Drug regimen	7469	231 (3%)		1a	Replaced by mode

4.2.7.2. Resolving Inconsistencies

The two possible causes for the discrepancies detected in the fields of selected attributes are human error in data entry and the design of the values of attributes of the database with no predefined values. The problem associated with existence of inconsistencies is that they reduce the quality of the final model and makes learning difficult for the algorithms (Han and Kamber, 2006).

Discrepancies were detected while extracting statistical summaries of attribute values. There are invalid values entered in the database. For instance under the field 'Occupation', the terms "no work", "job less", "Jobless", "No worker" etc were used to describe people who don't have jobs. Therefore for the sake of consistency, the researcher corrected them as one category "Unemployed". There are also other cases of expressing an occupation by different words or spellings (correct and erroneous). Therefore the researcher had to choose a single term that can serve instead of them.

4.2.7.3. Handling Outliers

A Database may contain data objects that do not comply with the general behavior or model of the data. These objects are considered as outliers. Deviation - based methods identify outliers by examining differences in the main characteristics of objects in a group. The degree to which numeric data tend to spread is called the dispersion, or variance of the data. The most common measures of data dispersion are range, based on quartiles, the inter quartile range and the standard deviation. Box plots can be plotted based on the five number summaries and are a useful tool for identifying outliers (Han and kamber, 2006).

Accordingly, the outlier values within the attribute of the dataset used for the research especially for numeric type attributes were explored and approached based on recommendations from different data mining literatures to handle the outlier values.

As stated in Han and kamber (2006) a common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5* IQR above the third quartiles or below the first quartile. In other words it is to mean that the values outside the limits:

Q3+(1.5*IQR) and Q1-(1.5*IQR) will be considered outliers values. Based on the recommendations Age, Weight and Baseline CD4 which are numeric in data type, have a kind of outlier us stated under section Table 4.2.

4.2.8. Data Reduction

4.2.8.1. Discretization

According to Chackrabarti et al.(2009), data discretization techniques are used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. This leads to a concise, easy-to-use, knowledge-level representation of mining results. Among the available data discretization techniques", binning has been selected to discretize the numeric attributes in to ranges of values for the independent variables and in to classes for the dependent variables.

Binning is a top-down splitting technique based on a specified number of bins (Chackrabarti, et al., 2009). The attribute values can be discretized by applying equal-width or equal-frequency binning. Accordingly Weka were used to bin weight.

a. Discretizing the Values of Age Attribute:

In order to reduce and simplify the original data, the researcher replaced numerous values of a continuous attribute by a smaller number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining result. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with high level concepts for instance Child hood, adolescent ,adult and old (Han and Kamber,2006). Here binning can be used to scale the values of "Age" attribute. But, there is already a meaningful discretized values used in this research are grouping the age value in to eight distinct value including "0-5","6-10", "11-15" "16-20", "21-25", "26-35", "36-45", and "46-53". But the researcher considered outliers in this already binned age value. By taking this in to consideration the researcher decided to take "8-15" , "16-20", "21-25", "26-35", "36-45", and "46-52.5". Therefore, 52.5 years is the upper limit for outliers so that age values beyond 52.5 years were treated as outliers. For the lower limit is 25-(1.5*IQR) which is 28-16.5=8.5. So Age values bellow 8.5 years can be considered outliers in this dataset. These grouped six distinct values are used to see patterns in "Age" if any. Therefore, the actual low level values are manually replaced by the corresponding high level grouped values.

Table 4.4: discretized value of age

No	Age attribute value	New value of age
1	8-15	A1
2	16-20	A2
3	21-25	A3
4	26-35	A4
5	36-45	A5
6	46-53	A6

As shown in the above table 4.4 range of age attribute value reduced in to six distinct high conceptual levels.

b. Discretizing the Values of Baseline CD4 Attribute:

Baseline CD4 count is classified according to the WHO cutoff point (200) to initiate ART in resource limited settings. With this regard, the continuous valued baseline CD4 counts are categorized in to two as "Below 200" and "Above 200" for the values less than 200 and greater than or equal to 200 respectively. But due to the large number of instances categorized under the category below 200 (86%), instances under this category are further classified in to three as "Below 50", "50 - 99", "100 – 150" and "151-199". This categorization is done based on a recommendation of experts from Jimma University specialized hospital specialists. Moreover, a national study conducted by FHAPCO indicated that more than 80% of ART following patients initiated therapy below 200 (FHAPCO, 2009).

Table 4.4: discretized value of CD4 count

No	Baseline CD4 attribute	New Baseline CD4 Value
1	<50	CD4A
2	50-99	CD4B
3	100-150	CD4C
4	151-199	CD4D
5	>=200	CD4E

c. Discretizing the Values of Weight Attribute:

As Chackrabarti, et al., (2009) binning is a top-down splitting technique based on a specified number of bins. The attribute values can be discretized by applying equal-width or equal-frequency binning. The researcher in collaboration with domain area experts approved standard based binning to avoid the possible deviation from the already accepted standards.

Accordingly, “weight” is binned in to six groups with an equal bin width of five with difference only made to the final group (76-93). The rest data values between 1 and 75 are equally partitioned in to five groups to form a total of six bins.

Table 4.4: discretized value of weight attribute

No	Weight attribute	New weight Value
1	4-15	W1
2	16-30	W2
3	31-45	W3
4	46-60	W4
5	61-75	W5
6	76-83	W6

4.2.9. Data Transformation

According to Ian et.al (2006), data transformation involves transforming or consolidating the data to a form appropriate for mining. Data transformation usually involves data smoothing, generalization of data, normalization of data, aggregation of data, and attribute construction.

In this study, generalization of data was used in order to give synonym word for Alive and Dead values of Class attribute. This was done based on the recommendation of domain experts. This is because predicting a patient as DEAD is not ethical and it put bad effect on their psychology. Therefore, for experiments done the class value were changed from Alive to high survival rate and DEAD to low survival rate.

Table 4.5: class attributes transformation

Attribute Name	Old value	New value
Status	Alive	High survival rate
	DEAD	Low survival rate

4.2.10. Dataset Format

Weka needs data to be prepared in some formats and file types. The data sets provided to this software are prepared in a format that is acceptable for Weka software. Weka accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format.

In order to prepare the data in such format the records from the Microsoft excel database are saved as a Comma Delimited (CSV) file. Once all processing is completed and the file is converted to .csv format, WEKA either process the .csv format itself or a file in the form of Attribute Relation File Format (.arff). For this study the data is given to the software in .arff format.

CHAPTER FIVE

5. EXPERIMENTATION AND ANALYSIS OF RESULT

In this chapter, the researcher describes the techniques that have been used in developing a model to predict HIV/AIDS survival rate. This research incorporated the typical stages that characterize a data mining process. This study has been organized according to Hybrid model. Here the researcher discuss the experimentation process by relating the steps followed ,the choice made , the task accomplished , the result obtained, evaluation of the model and results , and present it in a way that the organization can easily understand and use it.

5.1.Experimental Setup

All experiments are done on the final dataset which has passed all the preprocessing operations to suit the selected mining algorithms. The dataset contains 7,700 records with twelve predicting and one outcome variables. Initially, the dataset was in Microsoft excel format and then converted in to CSV in order to be read by WEKA machine learning software. To ease the repeated access of the file, it has been converted in to Attribute Relation File Format (ARFF) using Weka. Moreover, these experiments are done by using of Weka, which is version 3.6. The big challenge here is that the outcome variables named with status contains imbalanced classes which can possibly deteriorate the classifier's predictive accuracy. Imbalanced class distribution is characterized as that there are many more instances of some classes than others. With imbalanced data, classification rules that predict the small classes tend to be fewer and weaker than those that predict the prevalent classes; consequently, test samples belonging to the small classes are misclassified more often than those belonging to the prevalent classes. Different

research indicate that, the imbalanced class distribution of a data set poses serious difficulty to most classifier learning algorithms which assume a relatively balanced distribution (Japkowicz and Stephen, 2002)

Therefore, the researcher used SMOTE automatic operations by filter where minority classes are oversampled by generating synthetic examples of minority class and adding them to the dataset. This way, the class distribution in the dataset changes and probability of correctly classifying minority class increases. The final dataset with its selected attributes before and after SMOTE is shown in Figure 5.1 below.

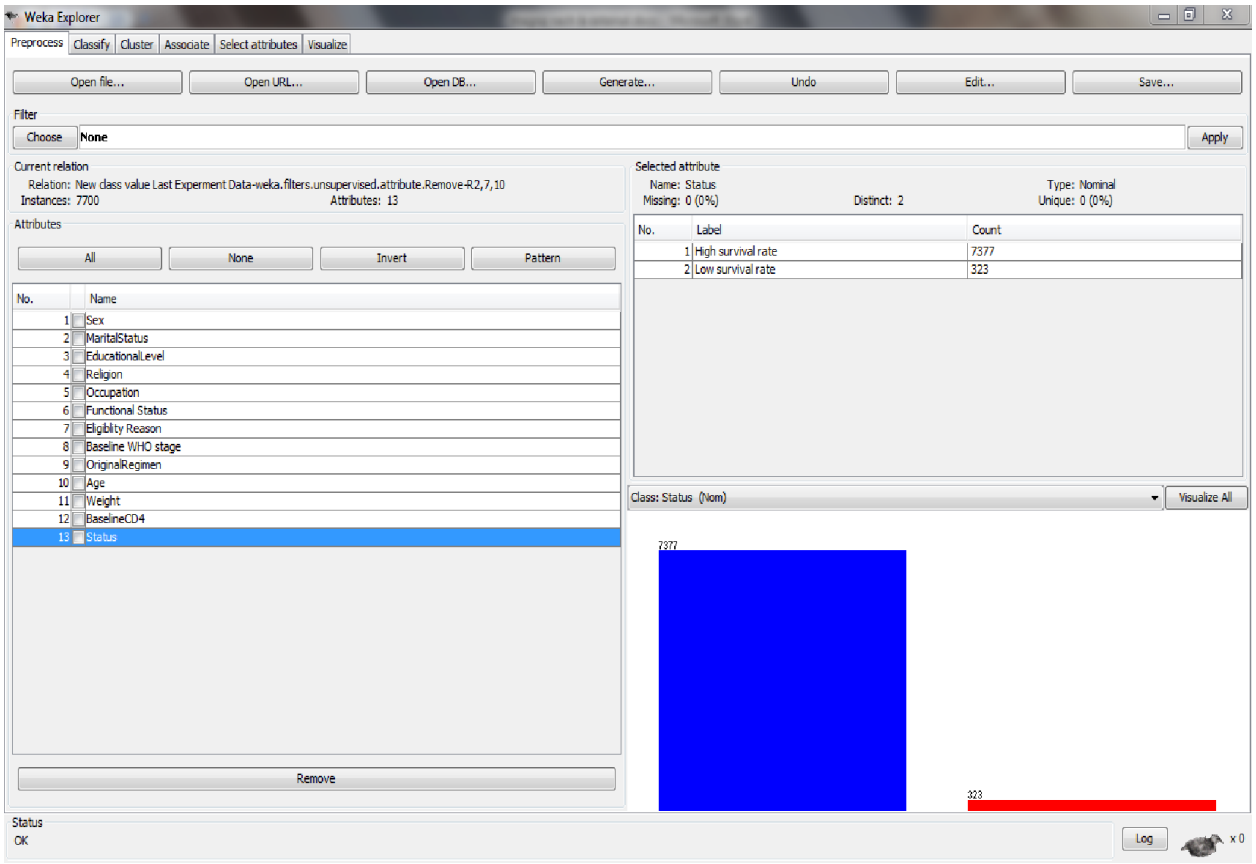

Figure 5.1a: Before SMOTE

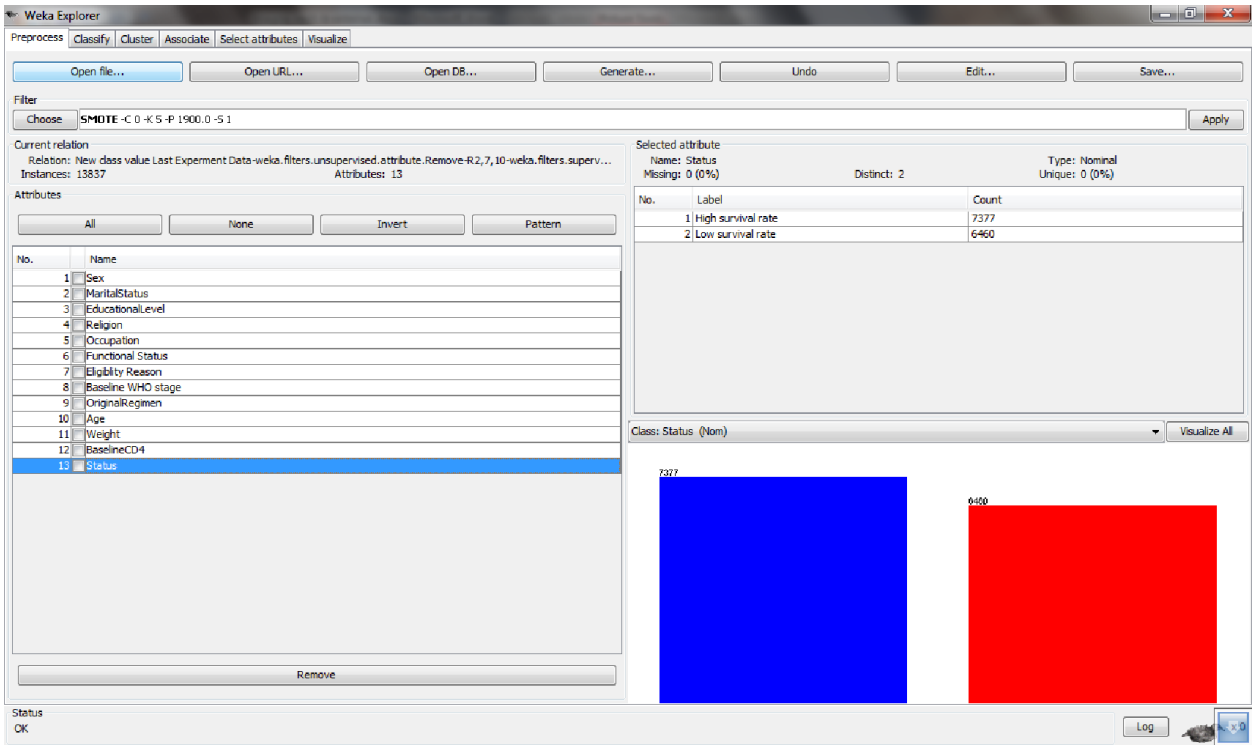

Figure 5.1b: After SMOTE

5.2.Model Building

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

5.2.1. Selecting Modeling Technique

Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research two classification techniques has been selected for model building. The analysis

was performed using WEKA environment. Due to easy of understanding and interpretation of the result of the model, the researcher selected PART and J48 for experimentation.

A PART algorithm is common types of rule induction technique which generate a model as a set of rules. In the meantime, the J48 algorithms of decision tree generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class output (Witten and Frank, 2005). J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA (Two Crows Corporation, 1999). It is WEKAs implementation of C4.5 top-down decision tree learners proposed by Quinlan (1986).

This research used C4.5 algorithms to predict survival rate of HIV/AIDS patients. This algorithm is implemented by modifying parameters such as confidence factor, pruning and unpruning, changing the generalized binary split decision classification and other option available. Therefore, it is very crucial to understand the available options to implement the algorithms, as it can make a significance difference in the quality of the result. In many cases, the default setting proved adequate, but to compare results /models and attain the research objectives other options are considered (Witten and Frank, 2005, Han and Kamber, 2006).

5.3.Attribute ordering

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

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

Where, Pi is the probability that an arbitrary tuples in sets of training data D belongs to certain class. Info (D) is also known as the entropy of D. After calculating information gain for each attribute, select the one with the highest information gain as the root node, and continue the calculation recursively until the data is completely classified by J48 algorithms in this case. For the purpose of this research, WEKA is used to compute the information gain and rank according to the importance of the attribute for the classifier. As a result of this, the following figure 5.2 depict ranked order of attribute based on their relevance for the reason that such attributes are very important for later experimentations by excluding the least relevant attributes.

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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 Status ):
        Information Gain Ranking Filter

Ranked attributes:
 0.503     3 EducationalLevel
 0.36413  11 Weight
 0.32412   7 Eligiblity Reason
 0.217    12 BaselineCD4
 0.1927    2 MaritalStatus
 0.18466   9 OriginalRegimen
 0.16571   4 Religion
 0.14494   8 Baseline WHO stage
 0.12475  10 Age
 0.12244   5 Occupation
 0.01714   6 Functional Status
 0.00172   1 Sex

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

Figure 5.2: summary of attribute ranking

As you can see from the result of attribute selection using entropy based information gain method of WEKA, out of 13 attributes, the top 10 the most relevant attributes are Educational Level, Weight, Eligibility reason, Baseline CD4, Marital Status, Drug Regimen, Religion, Baseline WHO stage, Age, and Occupation.

5.4.Running Experiments

For training and testing the classification model the researcher used two methods. The first method is percentage split method, where 70 % of the data used as training and the remaining 30% testing. The second method is K-fold cross validation methods, the data was divided into 10 folds, some fold is used as testing and the remaining folds are used as training.

Based on the above methods establishing scenario for model to be developed is very important to see the model result and analysis of each result; to compare the result of one model with the previous one and finally help us to find out the outperforming model based on criteria of evaluation. Consequently for both of the methods following scenario has been done for each of two selected model with default parameter value of WEKA 3.6 software.

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

5.4.1. Model building using J48 decision tree

In Hybrid methodology, data mining particularly model building is the crucial steps that should be dealt with. Model building is an iterative process's. Therefore in this study four different experiments were conducted altering the parameter using J48 algorithms and PART algorithms for building best predictive model.

Using aforementioned methods; the experiment has been conducted with two (2) scenarios for each method.

Table 5.1: Type of experiment

Type of Experiments	Scenarios
70/30 split criteria J48 Tree Model Generation	J48 unpruned with all attributes
	J48 pruned with all attributes
10 cross fold validation J48 Tree Model Generation	J48 unpruned with all attributes
	J48 pruned with all attributes

All of the 13 attributes, which are selected for model building, are fed as independent variables and the dependent variable of Status is a target class. The algorithms builds starting at each node it were sent either left or right according to some test. Eventually, it reached a leaf node and be given the label associated with that leaf. In general, this research was more interested in generating rules that best predict the survival rate and to come up with an understanding of the most important factors (variables).

J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Accordingly, based on two methods selected the following series of experiment has been conducted. Summary of experimentation with J48 algorithms result is depicted in Table 5.2

One of the compulsory steps of Hybrid methodology next to model building is evaluation of the model. Accordingly, the performance of the model has been evaluated based on the following criteria including performance accuracy, confusion matrix value, and True Positive rate and False Negative rate, Number of leaves and size of the tree generated and ROC curves and execution time.

As shown below in Table 5.2 the experimentation has performed in two methods (in 70-30 percentage split and 10 -fold cross validation test option respectively.) The first experiment has be tested with two scenario mentioned above with 70- 30 percentage split criteria. When we compare the result of method I. Scenario # 1 (pruning with all attribute) is the best model based on correctly classified instance (out of 4151, 4031 instances are correctly classified), performance registered (97.109% accuracy and TP rate of 97.1 % (classifying correctly high survival rate as high survival rate and low survival rate as a low survival rate). And also, in terms of minimum number of leaves and size of tree to build a models are pruning with all attribute is best model.

Table 5.2: Summary of J48 experiments

Model Characteristics	Experiment (Scenario)			
	1	2	1	2
Test option	Method I (70/30)		Method II (10 cross fold)	
Pruned	yes	no	yes	no
Attribute	All	All	All	All
Accuracy	97.109	95.88	97.477	96.7
Time Taken	0.49	0.65	0.2	0.4
No of values	45	852	45	852
Size of trees	55	1058	55	1058
AV.TPR (%)	97.1	95.9	97.5	0.966
AV.FPR (%)	63.3	4.3	2.8	0.036
AV.PR (%)	97.2	95.9	97.6	0.966
AV.RR (%)	97.1	95.9	97.5	0.966
AV.ROC (%)	97.4	95.6	97.7	0.966
CCI	4031	3980	13488	13362
ICI	120	171	349	475

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

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

The last compares in method one is made between Average ROC curve rate which registered a performance of 97.4% and 95.6% in pruned and unpruned respectively.

As a result Scenario # 1 (Building decision tree pruning with all attribute) is selected. But, in the second experiment or method two (10-fold cross validation cases) two of the stated trial was conducted. When we compare the result of each scenario developed model, scenario # 1 (building decision tree pruned with all attribute) is best on accuracy which registered 97.4 % and correctly classified instance which is accounts 13488 out of 13837 . Furthermore, based on criteria of minimum time taken for building model, scenario # 1 (building model with prunned with all attribute) is also best. And also, the number of tree and size of tree for scenario # 1 is minimums so it is best because it reduced the complexity of the generated tree. The average ROC curve performance measure indicates scenario #1 had 97.7 % performance and scenario #2 had 96.6%. This indicates that scenario #1 average ROC performance is better than scenario #2.Finally methods two scenario number one better than other model in terms of accuracy, correctly classified instance, time and ROC curve.

Accordingly, scenario #1 (Building pruned decision tree with all attribute) of 10- fold cross validation (method II) selected as the best J48 decision tree model. See the details confusion matrix in Appendix 3.

5.4.2. Comparison of Method I and Method II test option

In general when we compare both methods in terms of classification model accuracy 10 fold cross validation (method II) is better than 70-30 percentage split (Method I) .The accuracy ranges from 97.1 to 97.4 but 70-30 percentage split 95.88 to 97.109. Additionally, WEKA experimenter tool was used to compare of the J48 model developed with various scenarios.

However, by the same token aforementioned, the researcher selected scenario # 1 (building decision tree pruned with all attribute) of 10 fold cross validation which registered 97.4% accuracy as a final J48 decision tree model for training and testing the dataset and validated the performance of the model.

The detailed confusion matrix as a result of pruned J48 algorithm with all attribute has been discussed below in Table 5.3

Table 5.3: Summary of confusion matrix for pruned J48 decision tree

		Predicted Survival Rate Status		
		High survival rate	Low survival rate	Total
Survival status (Actual)	High survival rate	7350	27	7377
	Low survival rate	322	6138	6460
	Total	**7672**	**6165**	**13837**

The base of this research is determining the attributes more important to predict the survival rate. In this respect, the J48 algorithms classify how those instances are correctly and incorrectly classified in labeled class.

So as indicated in the above Table 5.3 based on the 10–fold cross validation test option the J48 learning algorithm scored an accuracy of 97.4%. This result shows that out of the total training

datasets 13488 (97.47 %.) records are correctly classified, while only 349 (2.5 %) of the records are incorrectly classified.

Besides, out of the total 7377 actually High survival rate HIV/AIDS patients, 7231 (99.63%) clients are correctly classified as High survival rate and the rest are misclassified as Low survival rate. And out of the total actually Low survival rate HIV/AIDS patients, 95.01% of clients are classified as Low survival rate and the rest are misclassified as High survival rate. This means the model has better performance in terms of correctly classifying High survival rate HIV/AIDS patients than Low survival rate.

Furthermore, evaluating the model based on sensitivity and specificity are very significance for decision making. For that reason .the result of the above confusion matrix indicate that the sensitivity of this test was (7350/7377) = 99.63 % and the specificity was (6138/6460) = 95.01%. The test indicates that the models appear to be pretty good. Because, based on the evaluation criteria, the classifier correctly classifies clients as High survival rate that has actually High survival rate with 99.63% accuracy.

The detailed confusion matrix as a result of pruned J48 algorithm with 70-30 has been discussed below in Table 5.4

Table 5.4 Summary of pruned J48 algorithms with 70/30 confusion matrix

		Predicted Survival Rate Status		
		High survival rate	Low survival rate	Total
Survival status (Actual)	High survival rate	2218	8	2226
	Low survival rate	112	1813	1925
	Total	**2330**	**1821**	**4151**

As we can see from Table 5.4, the resulting confusion matrix shows that the J48 decision tree algorithm scored 97.1% accuracy. That means out of the total 4151 records 4031 are correctly classified and the remaining are misclassified. In other words, the confusion matrix of this experiment has shown that 2218 (99.64%) of the 2226 total High survival rate patients are correctly classified as Low survival rate. But 23 (0.3 %) of them are misclassified as Low survival rate.

5.4.3. Model building using PART Rule Induction Algorithms

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

To build the Rule induction model, WEKA software package and the same 8023 ART dataset was used as an input respectively. The experiment was performed analogously as the researcher did in former model. That means the experiment was divided in two methods as the first method is with 70-30 percentage split criteria evaluate the two scenarios and Method two as experiment

with 10-fold cross validation for two scenarios again. The parameters are partially adjusted and default value was used with all attribute. Accordingly, the experiment of all scenarios with methods one was illustrated below in Table 5.5

Table 5.5: Experiment results of PART algorithms for two methods

Model Characteristics	Experiment (Scenario)			
	1	2	1	2
Test option	Method I (70/30)		Method II (10 cross fold)	
Pruned	yes	no	yes	no
Attribute	13	13	13	13
Accuracy	96.89	95.2	97.1	95.48
Time Taken	1.23	1.92	0.42	2.01
No of Rules	92	558	92	558
AV.TPR (%)	96.9	95.2	97.1	95.5
AV.FPR (%)	3.4	4.8	3.2	4.5
AV.PR (%)	97	95.2	97.2	95.5
AV.RR (%)	96.9	95.2	97.1	95.5
AV.ROC (%)	98.1	95.8	98	95.9
CCI	4022	3953	13436	13212
ICI	129	198	401	625

Key: The keys have the same meaning as explained in J48 for table 5.2

As shown in the above table 5.5, the registered performance is better that induction rule learner in case of pruned instead of unpruned one. And also, in case of J48 decision tree classifier, the researcher selected the one which register best with pruned case too.

Consequently, among the two scenarios experimented with 70-30 percentage split (method I) with default parameters, scenario #1 registered better performance of 96.89%. This shows that out of the training set of 4151 records 4022 (96.89%) of the records are correctly classified, while 135 (3.2 %) of the records are misclassified.

Table 5.6: Confusion Matrix of pruned PART Model (method II)

		Predicted Survival Rate Status		
		High survival rate	Low survival rate	Total
Survival status (Actual)	High survival rate	2205	27	2232
	Low survival rate	108	1817	1925
	Total	**2313**	**1844**	**4157**

Furthermore, the resulting confusion matrix also shown in Table 5.6 that out of the total 2232 High survival rate patients 2205 (98.79%) of them are correctly classified in their respective class, while 27 (1.2%) of the records are incorrectly classified in the Low survival rate classes or wrongly classified patients as Low survival rate. In other hand out of the total number 1925 of Low survival rate patients 1817 (94.38%) of them are correctly classified as Low survival rate but 108 (5.61%) of the patient records are misclassified.

Generally, from the above experiment for the two classes, classifying patients as High survival rate outweigh than Low survival rate patients.

The second method employed in this research to experiments with 10 fold cross validation test options. The result of the two scenario experiment was displayed in Table 5.5. In the same way, the experiment was made with different scenario for same test option and registered better performance in case of scenario #1 which accounts 97.1% of accuracy than other model. In this method of experiment, the PART algorithm brings better performance in case of the size of tree or rules than in unpruned ones, because, it reduce the complexity and increase easily understanding of the rules.

As shown in table 5.6 below the test result of the confusion matrix the model developed by the PART Algorithm with the pruned 10- fold cross validation, the model scored an accuracy of

97.1%. This shows that from the total 13837 test data, 13400 (97.1%). of the records are correctly classified, while 401 (2.8%) of them are misclassified.

Table 5.6: Confusion matrix of pruned PART algorithms with 10 folds crosses validation

		Predicted Survival Rate Status		
		High survival rate	Low survival rate	Total
Survival status (Actual)	High survival rate	7295	82	7377
	Low survival rate	319	6141	6460
	Total	**7614**	**6223**	**13837**

Additionally, you can also observe that out of the total 7377 actual High survival rate patient records 7295 (98.88%) of the records are correctly classified as High survival rate patients, while 82 (1.1%) of them are incorrectly classified in the Low survival rate class. Also out of 6460 total Low survival rate patients 6141 (95.06%) of the records are correctly classified as Low survival rate and the rest 319 (4.9%) are misclassified. when we compare the High survival rate and Low survival rate class, classifying clients as High survival rate correctly perform better than Low survival rate.

To sum up, in the above experiment J48 decision tree classifier (method II scenario #1 with 10-fold cross validation test option) that is conducted using default parameters values generates a better classification model with a better classification accuracy than 70-30 percentage split test options. So again J48 Experiment scenario #1 of 10.fold cross validation model selected as the best model of J48 for classifying patient's records as High survival rate and Low survival rate.

5.5.Comparison of J48 and PART models

Selecting a better classification technique for building a model, which performs best in handling the prediction and identifying significant attribute of survival rate of patient is one of the aims of this study. For that reason, the two selected classification model with respective best performance accuracy are listed in table 5.7

Table 5.7 comparison of J48 and PART models

Algorithms	Performance (%)	Time taken in (Sec.)	Correctly Classified	Misclassified
J48	97.4%	0.2	13488 out of 13837	349
PART	97.1%	1.2	13436 out of 13837	401
10 fold cross validation test option				

The result showed in table 5.7 that J48 decision tree outperforms PART rule induction by 0.3%. This means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular survival rate of HIV/AIDS patients. So, the model that is developed with the J48 algorithm classification technique is taken as the final working classification model to determine or predict survival rate of HIV/AIDS patients.

5.6.Generating Rules from Decision Tree

After consecutive experiments in building the best decision tree model, the next step were to generate rules by tracing through the branches up to leafs. A rule is a correlation found between the main variable (dependent) and the others (independent). The corresponding rules extracted form decision trees is listed below (see Appendix 5) and some of the rules believe to be interesting are randomly selected and presented as follows.

1) If Educational Level = Primary AND BaselineCD4= 100-150 AND Eligibility Reason = CD4 AND Baseline WHO stage = WHOStage2 AND Functional Status = Working: Then High survival rate (359.0)
2) If Educational Level = Tertiary AND BaselineCD4= 100-150 AND Baseline WHO stage = WHOStage2 AND Marital Status = Married: Then High survival rate(97.0)
3) If Educational Level = Primary AND Drug Regimen = 1e AND Age = 16-20: Then High survival rate(82.0)
4) If Weight = 16-30 AND Functional Status = Working AND BaselineCD4= 100-150 AND Drug Regimen = 1a: Then High survival rate(43.0/1.0)
5) If Weight = 46-60 AND BaselineCD4= 100-150 AND Baseline WHO stage = WHOStage3 AND Eligibility Reason = CD4 AND Educational Level = Secondary: Then High survival rate(205.0)
6) If Educational Level = Primary AND Drug Regimen = 1e AND Baseline WHO stage = WHOStage1: Then High survival rate (57.0)
7) If Educational Level = Primary AND Occupation = Un employed AND Baseline WHO stage = WHOStage1 AND Drug Regimen = 1a: Then High survival rate(34.0)
8) If Eligibility Reason = CD4 AND Educational Level = No Education Level AND Religion = Orthodox AND Occupation = self employed AND Age = A4 AND Functional Status = Working: THEN Low survival rate(38.0)
9) If Functional Status = Bedridden AND Drug Regimen = 1a AND Eligibility Reason = CD4 AND Weight = 31-45: THEN Low survival rate (39.0)
10) If Eligibility Reason = CD4 AND Educational Level = No Education Level AND Drug Regimen = 1a AND Baseline WHO stage = WHOStage4: then Low survival rate (342.0)
11) If BaselineCD4 = 151-199 AND Baseline WHO stage = WHOStage2 AND Educational Level = Secondary AND Marital Status = Married: then High survival rate (31.0/2.0)

For instance Rule 11 interpreted as clients whose baseline CD4 is between 151 and 199 and baseline WHO stage is stage 2 and his educational level is secondary and his marital status is married then the classifier grouped to high survival rate (31.0/2.0). The number in the brace indicates that in the data set 31 records that are exactly satisfied this rule and 2 records indicate misclassified to these rules.

Similarly, the remaining rules can be interpreted in the above way.

5.7.Discussions of Results

The rules generated from the decision tree in section 5.6 predict the survival rate of HIV/AIDS patients. The rule considers different conditions of the attributes age, Sex, Education level, drug regimen, eligibility reason, Baseline CD4, Baseline WHO stage and functional status to predict for the class value Status. The attributes age, Education level, drug regimen, eligibility reason, Baseline WHO stage and Baseline CD4 were identified as having a higher statistical significance in classifying the predicted value for Status of patients.

In the rules extracted, patients who had primary level of education and 1e drug regimen; and their reason for eligibility was CD4 with age gap between 16 and 20 had a good opportunity of having high survival rate. Whereas those with no educational background with a regimen of 1a and who are in stage 4 in the baseline of WHO, are expected to have low survival rate. As it is observed from rules 1 to 5 above, HIV/AIDS patients survival is associated with education level, baseline CD4, baseline WHO stage, age of the patient, marital status and drug regiment and even functional status.

This study identified attributes that significantly indicate the level of survival rate. Identifying patients at a low probability of survival has the advantage that due attention would be given to the risk group during their follow up to maximize the survival while they are taking ART.

This study found that the significant predictors of lower chance of survival in patients living with HIV/AIDS ON ART were: older age, low CD4 count at baseline, no education level, and low weight at baseline.

The functional status of patients can be seen as an indicator of the severity of the patient's health condition. Those patients who are in ambulatory and working functional status have the strength to work and engage themselves in household activities which may help them generate additional income and improve their quality of life. Even though patients who are staying in bed in hospitals are accessing medical care and support, cannot do anything by their own to create a stress-free environment for them. Thus, as expected, patients with ambulatory and bedridden status are low survival rate than working patients.

Aging here is also associated as reason for the survival of HIV/AIDS patients. Those with age range below 40 have shown improvements in their survival rate as high.

Marital status and Educational level has great impact as a reason for the survival of the patients. Those who are married and better educated are associated with high probability of having High survival rate.

The overall model building process made by employing both decision tree and partial tree techniques demonstrated that data mining is a method that should be considered to support public health care prevention and control activities.

The model building process also revealed the significance of some variables that were not initially suspected to be very important. For example, variables like “Educational level”, “age”, and “marital status” were identified as important variables to predict survival rate among HIV/AIDS patients.

Although both partial tree and decision trees showed comparable accuracy and performance in predicting the survival rate, the decision tree approach seems more applicable and appropriate to the problem domain.

To determine the importance of the above rules and the attributes used to construct those rules, the association of the attributes with the predicted class predicted by rules was evaluated based up on comments given by domain experts and reports of previous research works.

5.8.Evaluation of the Discovered Knowledge

Rules generated from the classification algorithm are presented in the form of if-then statements so that domain area experts can use it by taking the values of the independent variables to predict survival rate of HIV patients easily. In order to reach the final goal, two data mining goals were set to guide the overall flow of the study. Accordingly, the first mining goal to be attained was: Given the socio-demographic data, baseline WHO clinical stage and baseline CD4 cells count, predict the survival rate of HIV patient’s i.e. high survival rate or Low survival rate. For instance, take a patient record with the following details; If Educational Level = secondary AND

Eligibility Reason = CD4 AND Drug Regimen = 1e AND BaselineCD4 = 100-150: then High survival rate (259.0). Taking the first rule obtained from the J48 algorithm rule list, the patient survival process was high.. This indicated that a patient beginning with the attribute level of education which is socio demographic data and other clinical parameters needs a due attention to reverse the low survival rate.

The second data mining goal was: From the identified predicting variables, determine those having a better prediction performance. In this study, all the variables are selected based on the comments collected from the domain area experts and a review made on related literatures done in the area. Accordingly, all of the twelve predicting variables except one "religion" are selected in the study. But during knowledge extraction some of the variables predominantly appeared in each of the rules while others occurred less frequently. Therefore, those which occurred frequently are taken as the most predicting variables than the others. Seven variables are observed to be much more important in the identified if-then rules. These are; Education level, weight, marital status, Baseline CD4, Eligibility reason, Drug regimen and Baseline WHO stage, are used many times, so that these variables have high predicting capability than the other five attributes.

5.9.Prototype development

The final objective of this study was developing a prototype interface that assists physician easy access to the identified knowledgebase. The final selected if-then rules are used to implement the selected best models. The programming tool used to host the identified rules is Microsoft visual basic 6.0. Therefore, only eleven rules which are suggested to be important by domain experts are placed in to this prototype which means all the rules for predicting survival rate of a patient

can't be answered by this prototype. The following picture is the main graphical user interface used to run the commands to predict survival rate of HIV/AIDS patients at high survival rate or low survival rate.

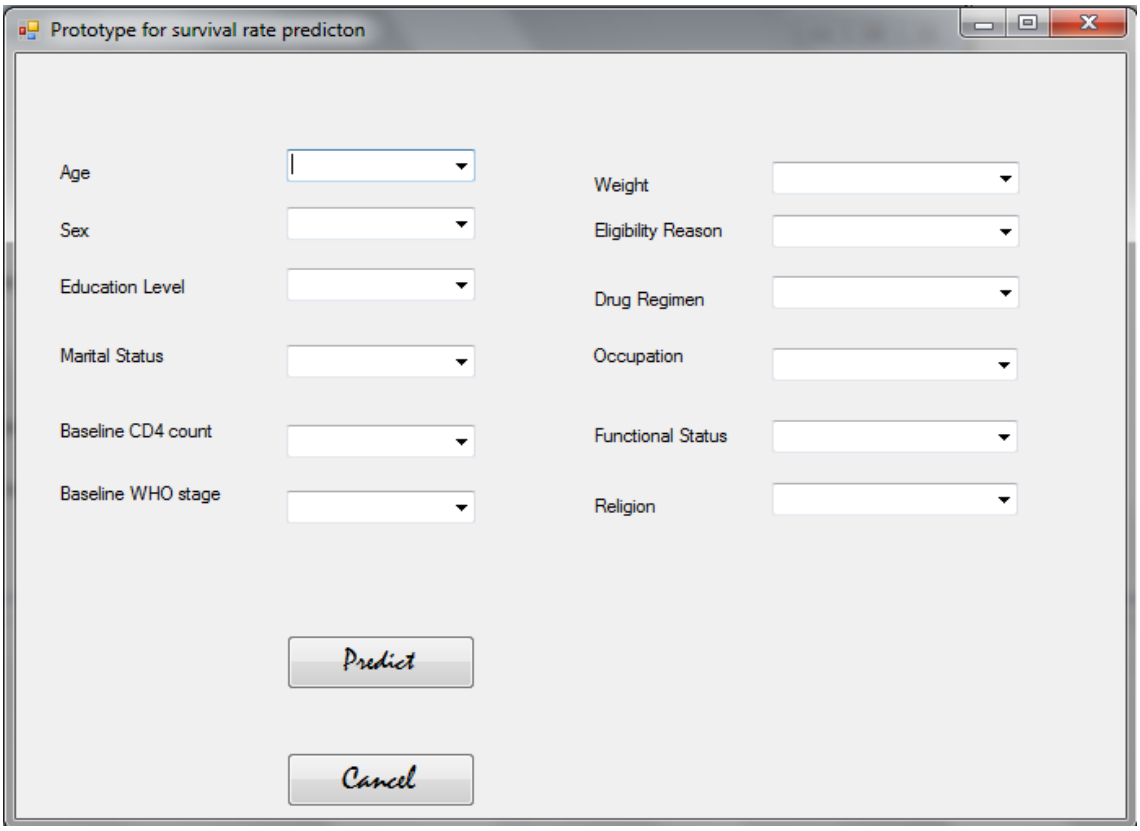

Figure 5.3 prototypes graphic user interface (GUI)

CHAPTER SIX

6. CONCLUSION AND RECOMMENDATIONS

6.1.Conclusion

A major challenge facing healthcare organizations (Antiretroviral therapy center, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable (Srinivas et al., 2010).

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

In this research, an attempt has been made to apply the data mining technology in support of identifying and predicting survival rate of HIV/AIDS patients using ART dataset. Data mining technology basically follows an iterative process consists of: Business understanding, Data understanding, Data collection, Data preparation, model building and evaluation. As discussed Hybrid data mining methodology in chapter three, there are six steps to above list for instance. The iterative nature of the process assist the data miner to back and forth at different and fix it where problem arise.

Accordingly, this research used and followed the six steps Cios et al (2007) methodology to investigate the stated problem. The methodology has strictly been followed while undertaking the experimentation. The data used in this research has been gathered from Jimma University specialized hospital. Once the data has been collected (9200 dataset), it has been preprocessed and prepared in a format suitable for the DM tasks. Data preparation took considerable time of the study.

The study was conducted using two classification techniques namely decision tree and rule induction. For model building and experimentation J48 and PART algorithms are used. Experimentation was conducted using two scenarios in two methods for each algorithm. By changing the training test options and the default parameter values of the algorithm, these models are tested and evaluated. As a result J48 algorithms was better performing with an accuracy of 97.4 % running with 10-fold cross validation with default parameter using 13 attribute than any experimentation done for this research purpose.

According to the result of J48 algorithms, all the 13 attributes are used to construct the decision rule set. Besides, using WEKA feature attribute selection is tried and tested the performance is 77 %. This implies that decrease in performance resulted because of missing of relevant attribute. To this end, with extensive discussion of domain expert and "educated guesses" out of the 13 attribute, only the following attributes are selected as the significant attributes of HIV/AIDS patients reason for the survival rate has been identified. These are Education Level, weight of patients, marital status, Baseline CD4 count, Eligibility reason, Drug regimen, Baseline WHO stage, age.

Analogously, PART rule induction algorithms were also register an encouraging performance of classification accuracy. This model correctly classify new instance of ART dataset as 97.1 % and 95.8 % with 10 fold cross validation with all attribute and default parameter. When we compare the result of two models developed based on accuracy registered to correctly classify new instance of ART data, J48 perform better. In general, in this study only few experiments are carried out using J48, and PART with few parameter setting. Missing values are simply replaced by most frequent values using WEKA 3.6 and manually for few attribute in consultation of domain expert which may sometimes create bias so that if time allows it is better to use other mechanism like filling extensively by consultation of domain experts. Furthermore, the researcher faced challenging in experimentation of balancing of the unbalanced dataset and getting of ART dataset. In spite of the challenges and weaknesses in data pre-processing and limited application of algorithms the results are encouraging. The results obtained from this research indicate that data mining is useful in bringing relevant information to the service providers as well as decision makers.

6.2.Recommendations

In this research work an attempt has been made to find out the potential applicability of data mining technology to support the survival of HIV/AIDS patients at Jimma University community and surrounding based on the data accumulated on ART. Even though, this research has been done for the academic exercise its results are found promising to be applied in addressing practical problems in prolonging life of HIV /AIDS patients.

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

This research has proven the applicability of J48 algorithms which automatically discover hidden knowledge that are interesting and accepted by domain expert. However, this research output doesn't integrate with knowledge based system. To do so, the researcher recommends implementing the discovered classification rules with domain knowledge as a knowledge based system that could be helpful for health professional, epidemiologists and other researchers who are conducting HIV/AIDS care, control and prevention.

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

For this work, balanced dataset is used to determine the survival rate of HIV/AIDS patients. However, researches can also be conducted using actual imbalanced data and other data mining technique such as Class Confidence Proportion Decision Tree (CCPDT). This is a robust decision tree algorithm for imbalanced datasets.

This research has been attempted to determine survival rate of HIV/AIDS patients with limited data set and 13 attributes collected from Jimma University specialized hospital. Further researches can also be conducted by increasing the number of attributes and datasets from all over regions in Ethiopia.

REFERENCE

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

AIDS Resource Center, Johns Hopkins University, Bloomberg School of Public Health, Center for Communication Programs (2005), Ethiopian National ART Strategic Communication Framework, website: http://www.etharc.org

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

Binyam T. (2008). Impact of malnutrition in survival of HIV infected children afterinitiation of ART, Unpublished M.Sc. Thesis Addis Ababa University.

Birru, A. (2009). Application of data mining techniques to support VCT for HIV: the case of center for disease controls and prevention (CDC).M.Sc. thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Bolton-Moore, M. (2007), 'Clinical outcomes and CD4 cell response in children receiving antiretroviral therapy at primary health care facilities in Zambia', Journal of the American Medical Association, vol. 298, no. 16, pp. 1888–1899.

Brady, M.T., Oleske, J.M., Williams, P.L., Elgie, C., Mofenson, L.M., Dankner, W.M., VanDyke, R.B. (2010). Declines in Mortality Rates and Changes in Causes of Death in HIV-1-Infected Children During the HAART Era', Acquired Immune Defic Syndr. for the Pediatric AIDS Clinical Trials Group219/219C Team 2010,';53(1):86-94

Bramer, M. 2007. Principles of Data Mining, Springer publisher, London, United Kingdom.

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

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

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

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

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

Factsheet on HIV and its treatment, (2010). Available from : http://aidsinfo.nih.gov/guidelines

Fayyad,U., Pazzani, M.J., Smyth, P., and Piatetsky, S. (1996). "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA

FHAPCO. (2007). Guidelines for paediatric HIV/AIDS care and treatment in Ethiopia.

FHAPCO. (2009). ART Scale-up in Ethiopia Success and Challenges.

Fontana, M., Zuin, G., Plebani, A. (1999). "Body composition in HIV infected children: relations with disease progression and survival", Am J clin Nutr,69:1282-86

Getu, S. (2007). Causes of road traffic accidents and possible counter measures on Addis Ababa-shashemene roads, Msc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Han, J and Kamber, M. (2006).Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

Ian et.al, (2010) "Data Mining: Practical Machine Learning Tools and Techniques" (3 Ed.) Elsevier.

Ian, W. (2005). Data mining: practical machine learning tools and techniques. Second Edition. San Francisco: Morgan Kaufmann Publishers.

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

Jiawei (2006). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

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

Larose, F. and Daniel T. Discovering Knowledge in Data - An Introduction to Data Mining. New Jersey, USA: John Wiley & Sons Inc.

Lior (2008).Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing Co. Pte.Ltd.

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

Mehmed, K. (2003). Data Mining: Concepts, Models, Methods, and Algorithms, ISBN13: 9780471228523, John Wiley & Sons Publisher 13.

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

Misganaw, T. (2005). Mining art data set to predict cd4 cells count: the case of Jimma, Bonga and Aman hospitals M.Sc. thesis, Addis Ababa University, Addis Ababa, Ethiopia.

National AIDS resource center (2009), Website http://www.etharc.org/resources/healthstat/nationalfactsheet/12-nationalfactsheet2009

National Factsheet, (2009). Website http://www.etharc.org/resources/healthstat/nationalfactsheet/13-nationalfactsheet2010

Nitesh (2012). Data Mining for Imbalanced Dataset: An Overview. Department of Computer Science and Enginering. University of Notre Dame: USA

Palella F.J., Delaney K .M., Moorman A.C., Loveless M.O., Fuhrer J., Satten G.A., Aschman D.J., Holmberg S.D. (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. N Eng. J Med.; 338:853-860.

Quinlane , J.R. (1986). Induction of Decision Trees. Kluwer Academic publisher. Boston (1),

Santos M. & Azevedo C (2005). "Data Mining" – Descoberta de Conhecimento em Bases de Dados. FCA Publisher

SAS Institute (1999). SAS Enterprise Miner – SEMMA [Internet]. Available from: http://www.sas.com/technologies/analytics/datamining/miner/semma.html

Selamawit, E. (2009). First line Antiretroviral Treatment Failure and Factors Associated with it in Addis Ababa, Master's thesis published in Addis Ababa University

Seyoum, Mekonen Y , Kassa A, Eltom A , DamitewT,Lera M, Felema B, Assefa Y (2009). ART scal-Up in Ethiopia: Success and Challenges, HAPCO, Plan, Monitoring &Evaluation Directorate

Shegaw A. (2002). Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project (BRHP). M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia;

Song R. J. (2007).'Efficacy of highly active antiretroviral therapy in HIV-1-infected children in Kenya', Pediatrics, vol. 120, no. 4, pp. 856–861.

Steven, J.R., Thomas, C.Q., Chris, B., Robert, C.B. (2003). Antiretroviral therapy where resources are limited. N Engl J Med; 348:1806-09.

Tekelu, T. (2010). Application of Data mining Technology on Antiretroviral Therapy(ART) data: a case of Adama and Assella Hospital. Master Thesis .Addis Ababa University. School of Information Science

Teklu, U. (2010). Application of data mining techniques on Antiretroviral Therapy (ART) data: the case of Adama and Asella hospitals. M.Sc. thesis, Addis Ababa University, Addis Ababa, Ethiopia

The CRISP-DM consortium, (2000). Step-by-step data mining guide available at: URL: http://www.crisp-dm.org/CRISPWP-0800.pdf .

Two crows Corporation (1999), Introduction to data mining and knowledge discovery (2nd Ed).by Edelstein, H., A. Potomac, MD: Two Crows Corp.

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

UNAIDS (2011).World AIDS Day Report of 2011. Geneva, UNAIDS.Retrived From: www.unaids.org/.../unaids/.../unaidspublication/2011/JC2216 Accessed on February 2014.

UNAIDS/WHO (2010). AIDS epidemic update UNAIDS 20 avenues Appia CH-1211 Geneva 27 Switzerland.

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

Virco 2008, HIV resistance learning system, module: Overview of HIV/AIDS, Virci Lab, New Jersey

Whitten I.H and Frank E. (2005). Data Mining: practical machine learning tools and techniques with java implementations. Morgan Kaufmann publishers. San Francisco.

WHO (2003).Scaling-up antiretroviral therapy in resource-limited settings. Treatment guideline for a public health approach. 2003 revision.

WHO (2006). Antiretroviral Therapy for HIV infection in adult"s adolescents. 2nd Edition;.

Witten, I. H., & Frank, E. ,2000 ,"Data mining concepts" ,New York, Morgan-Kaufmann WHO and UNAIDS (2003). Treating 3 million by 2005: Makingit happen.

Witten, T. and Frank, E.(2000) Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco

World Health Organization, (2009). Hospital Care for Children: Guideline for the Management of Common Illnesses with Limited Resources, 20 Avenue Appia, 1211 Geneva 27, Switzerland.

Zhang X, (2007). HIV/AIDS Relative Survival Analysis, Master's thesis published in Georgia University.

APPENDIX 1: J48 Pruned tree 10 fold cross validation

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: class date Last Experment Data-Weka.filters.unsupervised.attribute.Remove-R2,7,10-Weka.filters.supervised.instance.SMOTE-C0-K5-P1900.0-S1

Instances: 13837

Attributes: 13

Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves : 45

Size of the tree : 55

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 13488 97.4778 %

Incorrectly Classified Instances 349 2.5222 %

Kappa statistic 0.9492

Mean absolute error 0.0478

Root mean squared error 0.1556

Relative absolute error 9.6016 %

Root relative squared error 31.187 %

Total Number of Instances 13837

=== Detailed Accuracy By Class ===

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

0.996 0.05 0.958 0.996 0.977 0.977 High survival rate

0.95 0.004 0.996 0.95 0.972 0.977 Low survival rate

Weighted Avg. 0.975 0.028 0.976 0.975 0.975 0.977

=== Confusion Matrix ===

a b<-- classified as

7350 27 | a = High survival rate

322 6138 | b = Low survival rate

Appendix 2: J48 unpruned tree 10 fold cross validation

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -U -M 2

Relation: class date Last Experment Data-Weka.filters.unsupervised.attribute.Remove-R2,7,10-Weka.filters.supervised.instance.SMOTE-C0-K5-P1900.0-S1

Instances: 13837

Attributes: 13

Test mode: 10-fold cross-validation

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

J48 unpruned tree

Number of Leaves : 852

Size of the tree : 1058

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 13362 96.5672 %

Incorrectly Classified Instances 475 3.4328 %

Kappa statistic 0.9309

Mean absolute error 0.0456

Root mean squared error 0.1811

Relative absolute error 9.1555 %

Root relative squared error 36.3091 %

Total Number of Instances 13837

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.977	0.048	0.959	0.977	0.968	0.966	High survival rate
0.952	0.023	0.973	0.952	0.963	0.966	Low survival rate
Weighted Avg. 0.966	0.036	0.966	0.966	0.966	0.966	

=== Confusion Matrix ===

a	b	<-- classified as
7209	168	a = High survival rate
307	6153	b = Low survival rate

Appendix 3: J48 Pruned tree 70/30 split criteria

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: class date Last Experment Data-Weka.filters.unsupervised.attribute.Remove-R2,7,10-Weka.filters.supervised.instance.SMOTE-C0-K5-P1900.0-S1

Instances: 13837

Attributes: 13

Test mode: split 70.0% train, remainder test

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

J48 pruned tree

Number of Leaves: 45

Size of the tree: 55

Time taken to build model: 0.1 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 4031 97.1091 %

Incorrectly Classified Instances 120 2.8909 %

Kappa statistic 0.9417

Mean absolute error 0.05

Root mean squared error 0.1661

Relative absolute error 10.0493 %

Root relative squared error 33.3051 %

Total Number of Instances 4151

=== Detailed Accuracy By Class ===

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

0.996 0.058 0.952 0.996 0.974 0.974 High survival rate

0.942 0.004 0.996 0.942 0.968 0.974 Low survival rate

Weighted Avg. 0.971 0.033 0.972 0.971 0.971 0.974

=== Confusion Matrix ===

a b<-- classified as

2218 8 | a = High survival rate

112 1813 | b = Low survival rate

Appendix 4:J48 unpruned tree 70/30 split criteria

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -U -M 2

Relation: class date Last Experment Data-Weka.filters.unsupervised.attribute.Remove-R2,7,10-Weka.filters.supervised.instance.SMOTE-C0-K5-P1900.0-S1

Instances: 13837

Attributes: 13

Test mode: split 70.0% train, remainder test

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

J48 unpruned tree

Number of Leaves : 852

Size of the tree : 1058

Time taken to build model: 0.09 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 3980 95.8805 %

Incorrectly Classified Instances 171 4.1195 %

Kappa statistic 0.9171

Mean absolute error 0.0487

Root mean squared error 0.1943

Relative absolute error 9.7794 %

Root relative squared error 38.9653 %

Total Number of Instances 4151

=== Detailed Accuracy By Class ===

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

0.969 0.054 0.954 0.969 0.962 0.956 High survival rate

0.946 0.031 0.964 0.946 0.955 0.956 low survival rate

Weighted Avg. 0.959 0.043 0.959 0.959 0.959 0.956

=== Confusion Matrix ===

a b<-- classified as

2158 68 | a = High survival rate

103 1822 | b = Low survival rate

Appendix 5: J48 pruned tree

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

EducationalLevel = Secondary: High survival rate (2018.0/121.0)

EducationalLevel = Primary: High survival rate (2501.0/92.0)

EducationalLevel = NoEducationLevel

| Religion = Orthodox

| | Weight = W1: High survival rate (76.0/2.0)

| | Weight = W3

| | | MaritalStatus = NeverMarried: High survival rate (44.0)

| | | MaritalStatus = Divorced: High survival rate (47.0/3.0)

| | | MaritalStatus = Married

| | | | BaselineCD4 = CD4A: High survival rate (3.0)

| | | | BaselineCD4 = CD4C

| | | | | Eligibility Reason = TCLandCD4: Low survival rate (4680.0/5.0)

| | | | | Eligibility Reason = CD4

| | | | | | Drug Regimen = 1e

| | | | | | | Baseline WHO stage = WHOStage1: Low survival rate (0.0)

| | | | | | | Baseline WHO stage = WHOStage2: High survival rate (25.0)

| | | | | | | Baseline WHO stage = WHOStage3

| | | | | | | | Functional Status = Working

| | | | | | | | | Occupation = Self employed: Low survival rate (59.0/2.0)

| | | | | | | | | Occupation = Un employed: Low survival rate (120.0/6.0)

| | | | | | | | | Occupation = employed: High survival rate (4.0)

| | | | | | | | Functional Status = Ambulatory: High survival rate (7.0)

| | | | | | | | Functional Status = Bedridden: High survival rate (2.0)

| | | | | | | | Functional Status = Appropriate: Low survival rate (0.0)

| | | | | | | | Functional Status = Delay: Low survival rate (0.0)

| | | | | | | | Functional Status = Regression: Low survival rate (0.0)

| | | | | | | Baseline WHO stage = WHOStage4: Low survival rate (0.0)

| | | | | | Drug Regimen = 1b: Low survival rate (0.0)

| | | | | | Drug Regimen = 1a: Low survival rate (1300.0/8.0)

| | | | | | Drug Regimen = 1c: Low survival rate (0.0)

| | | | | | Drug Regimen = 4c: Low survival rate (0.0)

| | | | | | Drug Regimen = 2c: Low survival rate (0.0)

| | | | | | Drug Regimen = 4b: Low survival rate (0.0)

| | | | | | Drug Regimen = 4a: Low survival rate (0.0)

| | | | | Eligibility Reason = TLC: High survival rate (11.0)

| | | | | Eligibility Reason = ClinicalOnly: Low survival rate (0.0)

| | | | | Eligibility Reason = ClinicalandCD4: Low survival rate (0.0)

| | | | | Eligibility Reason = Clinical,TCLandCD4: Low survival rate (0.0)

| | | | BaselineCD4 = CD4B: High survival rate (17.0/2.0)

| | | | BaselineCD4 = CD4D: High survival rate (7.0/1.0)

| | | | BaselineCD4 = CD4E: High survival rate (15.0/3.0)

| | | MaritalStatus = Separated: High survival rate (48.0/1.0)

| | | MaritalStatus = Widow/widower: High survival rate (51.0/1.0)

| | | MaritalStatus = Livingtogether: Low survival rate (0.0)

| | Weight = W4: High survival rate (309.0/8.0)

| | Weight = W5: High survival rate (16.0/1.0)

| | Weight = W6: High survival rate (3.0)

| | Weight = W2: High survival rate (6.0/1.0)

| Religion = Muslim: High survival rate (585.0/16.0)

| Religion = Protestant: High survival rate (122.0/3.0)

| Religion = other: High survival rate (1.0)

| Religion = Catholic: High survival rate (7.0)

EducationalLevel = Others: High survival rate (1148.0/38.0)

EducationalLevel = Tertiary: High survival rate (605.0/29.0)

Appendix 6: Decision rule list

1) If Educational Level = Primary AND Drug Regimen = 1e: then High survival rate (913.0/13.0)
2) If Weight = 16-30 AND Marital Status = Married: then High survival rate (117.0/2.0)
3) If Educational Level = Secondary AND BaselineCD4 = 100-150 AND Eligibility Reason = CD4: then High survival rate (315.0/3.0)
4) If Educational Level = No Education Level AND Marital Status = Married AND Weight = 31-45 AND Religion = Orthodox AND Eligibility Reason = TCLandCD4: then Low survival rate (4682.0/6.0)
5) If Eligibility Reason = TLC AND Weight = 46-60 AND Occupation = self employed AND BaselineCD4 >=200: then High survival rate (81.0/5.0)
6) If BaselineCD4 = 151-199 AND Baseline WHO stage = WHOStage2 AND Educational Level = Secondary AND Marital Status = Married: then High survival rate (31.0/2.0)
7) If Eligibility Reason = TLC AND Weight = 31-45 AND BaselineCD4 = 100-150 AND Drug Regimen = 1c: then High survival rate (45.0/2.0)
8) If Eligibility Reason = TLC AND Weight = 46-60 AND Drug Regimen = 1a AND Functional Status = Working AND Marital Status = Married AND Occupation = Self employed: then High survival rate (24.0/3.0)
9) If Eligibility Reason = CD4 AND Marital Status = Married AND Religion = Orthodox AND Educational Level = No Education Level AND Drug Regimen = 1a AND BaselineCD4 = 100-150 AND Baseline WHO stage = WHOStage3 AND Age = 26-35 AND Functional Status = Working: then Low survival rate (343.0/1.0)
10) If Eligibility Reason = CD4 AND Educational Level = No Education Level AND Drug Regimen = 1a AND BaselineCD4 = 100-150 AND Baseline WHO stage = WHOStage3 AND Sex = M: then Low survival rate (342.0)
11) If Eligibility Reason = CD4 AND Educational Level = No Education Level AND Baseline WHO stage = WHOStage4: then Low survival rate (342.0)
12) If Educational Level = No Education Level AND BaselineCD4 = 100-150 AND Baseline WHO stage = WHOStage3 AND Religion = Orthodox AND Occupation = Un employed AND Functional Status = Working: then Low survival rate (236.0/8.0)
13) If Educational Level = Primary AND Functional Status = Working AND Eligibility Reason = CD4 AND Marital Status = Married: then High survival rate (13.0)
14) If Educational Level = Secondary AND Baseline WHO stage = WHOStage2 AND Eligibility Reason = TLC: then High survival rate (15.0)
15) If BaselineCD4 >=200 AND Religion = Orthodox AND Functional Status = Working AND Weight = 46-60: then High survival rate (39.0/9.0)
16) If Educational Level = Secondary AND Baseline WHO stage = WHOStage1 AND Drug Regimen = 1a: then High survival rate (9.0)
17) If BaselineCD4 = 151-199 AND Baseline WHO stage = WHOStage3 AND Eligibility
18) If Eligibility Reason = CD4 AND Educational Level = No Education Level AND Occupation = Self employed AND Age = 16-20 AND Functional Status = Working AND Drug Regimen = 1a: then Low survival rate (38.0)

19) If Eligibility Reason = CD4 AND Educational Level = No Education Level AND Religion = Orthodox AND Occupation = Self employed AND Age = 26-35 AND Functional Status = Working: then Low survival rate (38.0)
20) If Functional Status = Bedridden AND Drug Regimen = 1a AND Eligibility Reason = CD4 AND Weight = 31-45: then Low survival rate (39.0)
21) If BaselineCD4 >=200 AND Functional Status = Working AND Age = 16-20: then High survival rate (9.0/1.0)
22) If Age = 16-20 AND Religion = Orthodox AND Occupation = Self employed AND Educational Level = No Education Level AND Functional Status = Ambulatory: then Low survival rate (20.0)
23) If Functional Status = Ambulatory: then Low survival rate (5.0)