ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT ANTIRETROVIRAL THERAPY INITATION TIME THE CASE OF ADAMA AND AMBO HOSPITALS, OROMIA REGIONAL STATE

BY

GETACHEW DEJENE

OCTOBER 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT ANTIRETROVIRAL THERAPY INITATION TIME THE CASE OF ADAMA AND AMBO HOSPITALS, OROMIA REGIONAL STATE

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

BY
GETACHEW DEJENE

OCTOBER 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT ANTIRETROVIRAL THERAPY INITATION TIME THE CASE OF ADAMA AND AMBO HOSPITALS, OROMIA REGIONAL STATE

BY

GETACHEW DEJENE

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.

_____

Getachew Dejene

October, 2013

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

Prof. Ahmed Ali

_____

Million Meshesha (Ph.D.)

_____

ACKNOWLEDGEMENT

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

Second and most importantly, I would like to acknowledge my advisors Professor Ahmed Ali and Dr. Million Meshesha for constructive comments, great deal of patience and guidance that they have provided me throughout the study.

My deepest gratitude also goes to my co-Advisor Dr. Wondwosen Amogne, from the School of Medicine for his support and comments on every of the problems I faced in this work.

I thank the staff in the Adama and Ambo ART Hospitals for their generous assistance, and friendship. I couldn't have started the study, if it was not for the interest and support contributed form Dr. Kebede and Ato Alemayehu. I am really indebted to Dr. Kebede for his support in evaluating the model/findings based on previously existing knowledge in the domain area; and to Ato Fiseha and W/ro Shito on dealing with problems in the data and technical facilitations.

I would also like to thank Addis Ababa University, School of Information Science and School of Public Health for financial support and overall facilitation of the research from the beginning until the end.

I would like to thank importantly the Health Informatics Program Coordinator Meseret Ayanaw for her continuous and valuable effort to realize my work. My classmates for their comments, constructive ideas and suggestions; to my best friends Girma Amare, Bereket Kidane and relatives for their moral support and understanding during the time I solely devoted to the study.

Finally I am also grateful to many others, who cannot be named here.

TABLES OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LISTS OF ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ANN	Artificial Neural Network
ART	Antiretroviral therapy
ARV	Antiretroviral
CART	Classification and Regression Trees
CD4	Clustered Differentiation 4
CDC	Center for Disease Control and Prevention
CRISP DM	Cross-Industry Standard Process for Data Mining
HCT	HIV Counseling and Testing
HIV	Human Immunodeficiency Virus
KDD	Knowledge Discovery in Data base
KDP	Knowledge Discovery Process
NGO	None Governmental Organization
ROC	Receiver Operating Characteristics
SQL	Structured Query Language
TLC	Total Lymphocytes Count
UNAIDS	United Nations AIDS
WHO	World Health Organization

ABSTRACT

Background: AIDS patients receive antiretroviral treatment (ART) which they need to take every day for the rest of their life. To maintain treatment efficacy, it is necessary to start the treatment at a suitable time. Although the debate regarding when to start antiretroviral therapy has been present for over two decades, consensus on this question has been hard to achieve. This lack of clarity continues in the current era, with major guidelines recommending very different treatment strategies.

Objective: The purposes of this research are to assess the applicability of different data mining techniques to predict the initiation time for Antiretroviral Treatment (ART), to identify attributes that are associated with initiation time of ART and to develop a model that can be used to predict the initiation time for Antiretroviral Treatment (ART) using data obtained from Adama and Ambo ART clinic.

Method: To undertake this study a hybrid Data mining process model has been employed. The study used 11,440 instances, ten predicting attributes and one outcome variables to run the experiments. Accordingly, Apriori algorithm is used to extract association rules while classification algorithms such as J48 decision tree, PART rule induction and Naïve Bayes were implemented to build predictive models.

Result: Experimental result shows that the model developed using AdaBoostM1withpruned PART registers the highest accuracy of 95.62% as compared to Naïve Bayes and J48. The finding of the study clearly presents that Sex, age, OACD4, OAWHO Stage, Family planning and Occupation attributes are best predicts used to predict ART Initiation Time.

Conclusion: The study comes up with a **predictive model** that assists practitioners to predict whether the pre-ART patients should start the treatment "immediately", “Early” or "Delayed".

CHAPTER ONE
INTRODUCTION

1.1. Background

1.1.1. HIV as Public health Challenge

Human Immunodeficiency Virus, well known by its acronym HIV, is the virus that causes Acquired Immunodeficiency Syndrome (AIDS) in humans. AIDS is the disease caused by HIV, which weakens the body's immune system until it can no longer fight off the simple infections that most healthy people's immune system can resist. Chronic fatigue, diarrhea, fever, weight loss, persistent cough, skin rashes, oral infections, swelling of the lymph nodes, and memory loss may be the early symptoms of AIDS. As the immune system becomes further compromised by HIV, opportunistic infections such as pneumonia, meningitis, cancers, and Tuberculosis (TB) can easily attack the body [1].

The emergence of the HIV epidemic is one of the biggest public health challenges the world has ever seen in recent history. In the last three decades HIV has spread rapidly and affected all sectors of society- young people and adults, men and women, and the rich and the poor. Sub-Saharan Africa is at the epicenter of the epidemic and continues to carry the full brunt of its health and socioeconomic impact [2].

The Acquired Immune Deficiency Syndrome (AIDS) disease has been one of the most destructive epidemics to hit Ethiopia. In 2006, there were 977,394 people living with the virus, and of these 258,264 require antiretroviral treatment (ART) [3]. Adults and adolescents account for 24% of antiretroviral (ARV) service coverage by December 2006 [1]. A Federal Ministry of Health (FMoH) report showed that 95,756 patients started ART treatment in Ethiopia. Out of these 71,773 (74.95%) patients are presently on ART [3]. HIV/AIDS is one of the key challenges for overall national development in Ethiopia. It has led to a seven-year loss in life expectancy, close to a million orphans and a loss of productivity and income at the workplace with severe effects at the household and community levels. The high rates of morbidity and mortality associated with HIV/AIDS have strongly affected the health sector and are among the major impediments to delivering quality care to its full capacity [4].

1.1.2. General Over view of ART

Antiretroviral therapy (ART) is defined as treating retroviral infections like HIV with drugs [4]. Antiretroviral therapy (ART) has been shown to be effective in slowing down the progression of AIDS and in reducing HIV-related illnesses and death [2]. Antiretroviral Therapy (ART) is treatment for AIDS that helps the body's immune system recover from the damage caused by infection with HIV. Although ART cannot cure AIDS, persons on ART will begin to feel better, eat more, and put on weight [2]. Their bodies will recover the ability to fight infections. As persons on ART treatment become well, they can care for their children and return to household activities and lead productive life, which benefits the household and national economies. They recover their sense of hope for the future and can become powerful advocates for prevention and mitigation of HIV in their families and communities. They may remain well for many years, but must continue to take Antiretroviral (ARVs) for the rest of their lives. Thus, ART is an important component of the global response to AIDS.

The standardization and simplification of treatment and monitoring continues to be the prime consideration underpinning WHO recommendations for the use of ART, in order to widen access to effective therapy in resource-limited settings where individualized patient management by physicians specialized in HIV medicine is not feasible [5]. Standardized clinical and, where available, immunological (CD4) evaluation to guide the initiation of ART, the use of appropriate formulations, and a symptom directed approach to monitor adverse events, are key to the simplified approach [5].

While the HIV prevalence in Ethiopia is relatively low for a Sub-Saharan country, which is at 1.5%, approximately, 800,000 people are living with HIV; almost 300,000 are currently on antiretroviral therapy. In 2011, almost 10 million people were tested for HIV and received their results [6]. The introduction and expansion of rapid diagnostic testing (RDT) for HIV has allowed for this increased access to diagnostic testing [6].

Adults and adolescents account for 24% of antiretroviral (ARV) service coverage by December 2006 [3]. A Federal Ministry of Health (FMoH) report showed that 95,756 patients started ART treatment in Ethiopia. Out of these 71,773 (74.95%) patients are presently on ART [3].

The WHO 2010 guidelines recommend that initiating HIV positive patients on antiretroviral therapy upon a CD4 count of 350 cells/μl or WHO stage III or IV. Providing CD4 testing throughout Ethiopia will be essential to initiating additional HIV positive patients on ART and improving overall patient health [6].

1.1.3 Eligibility for ART

The optimum time to commence ART is before patients become unwell or present with their first opportunistic infection. Immunological monitoring (CD4 testing) is the ideal way to approach this situation [7]. A baseline CD4 cell count not only guides the decision on when to initiate ART but is also essential if CD4 counts are to be used to monitor ART. The following table summarizes the immunological criteria for the initiation of ART [7].

Table 1.1:- CD4 criteria for the initiation of ART in adults and adolescents

CD4 (cells/mm30	**Treatment recommendation**
<200	Treat irrespective of clinical stage
200−350	Consider treatment and initiate before CD4 count drops below 200 cells/mm3
>350	Do not initiate treatment

The National antiretroviral initiation criteria for both adults and adolescents states that, in places where CD4 cell count is available [7].WHO clinical stage IV, irrespective of CD4 cell count, WHO stage III and if CD4 cell count is less than 350 and all WHO stage IV and CD4 cell count is less than 50 are criteria for eligible patients to start ART. However, for places where CD4 cell count is not available, WHO stage IV and WHO stage III, irrespective of TLC, WHO stage II and if TLC is less than or equal to 1200 are eligible to initiate ART treatment [3].

In developing countries where there is resource limitation using CD4 cell count to identify if a patient is eligible is difficult [7]. ART The current guidelines for Antiretroviral Therapy (ART) from the World Health Organization reflect the 2003 changes to the guidelines and recommend that in resource-limited settings, HIV –infected adults and adolescents should start ART when HIV infection has been confirmed and one of the following conditions is present [7].

1.1.4 Criteria for Clinical Staging

In 2006, the World Health Organization (WHO) released revised criteria for clinical staging of HIV disease in adults and adolescents. These criteria allow physicians in countries with poor resource setting to determine the appropriate time to begin antiretroviral treatment. In many areas of the World, physicians do not have access to labs where they can perform CD4 and viral load tests, which are used in developed countries to determine an individual's disease progression. According to WHO [7]. the following are the clinical stages of HIV progression.

Criteria for Stage I: During the first stage of HIV, an individual generally has flu like symptoms which last for a week or two. WHO provides the following criteria for placing a patient in this stage [7]:-

- Asymptomatic
- Persistent generalized lymphadenopathy (the swelling or enlargement of the lymph nodes).

Criteria for Stage II: In stage II, many people are completely asymptomatic, but others demonstrate a number of physical symptoms that healthcare providers can use to stage the patient. WHO criteria for this stage include the following:

- Moderate unexplained weight loss
- Recurring respiratory tract infections
- Herpes Zoster (shingles)
- Angular cheilitis (lesions at the corner of the mouth)
- Recurring oral ulceration
- Papular pruritic eruptions (skin rash possibly related to insect bites)
- Seborrhoeic dermatitis (a skin disorder that causes scaly, itchy, flaky skin)
- Fungal nail infections.

Criteria for Stage III: In stage III, HIV patients begin to exhibit more serious symptoms. This is also when opportunistic infections begin to take advantage of the weakened immune system. WHO criteria for placing a patient in this stage include the following:

- Unexplained severe weight loss

- Unexplained chronic diarrhea for longer than one month
- Unexplained persistent fever, either intermittent or constant
- Persistent oral candidiasis (yeast infection of the mouth)
- Oral hairy leukoplakia (a white patch on the side of the tongue with a hairy appearance)
- Pulmonary tuberculosis
- Severe bacterial infection (for example, pneumonia, meningitis, and empyema)
- Acute necrotizing ulcerative stomatitis (inflammation of the stomach mucous lining), gingivitis (inflammation of the gums), or periodontitis (inflammation of the tissue that supports the teeth)
- Unexplained anemia (lack of hemoglobin the blood cells), neutropenia (low number of a certain type of white blood cell called neutrophil), and/or chronic thrombocytopenia (low platelet count).

Criteria for Stage IV (AIDS): In stage IV, a patient is considered to have progressed from HIV to AIDS. This stage is characterized by more severe symptoms and an even greater number of opportunistic infections.

1.1.5 ART and its Practice in Ethiopia

In Ethiopia there are about 119 hospitals and 412 health centers and the country's health infrastructure has the potential to scale up access to antiretroviral therapy. But there is a substantial shortage of health workers to serve the needs of a rapidly expanding population. This shortage is aggravated by high turnover among health workers, especially physicians and counselors, throughout Ethiopia [5]. Antiretroviral therapy is provided only at referral and provincial hospitals. Scaling up antiretroviral therapy services would require an extension within the health system to include more peripheral facilities. Systems to procure and distribute drugs and surveillance, Information management system for managing patients' record and monitoring and evaluation systems also need to be strengthened [5].

Even though for ART eligibility criteria discussed in section 1.1.3 and 1.1.4 can be considered, from developing countries specifically Ethiopian context, there are also several concerns. The number of HIV positive persons enrolled into ART care in Ethiopia is nearly 3,000 patients per month [6]; however, CD4 testing services that are critical for treatment monitoring and initiation of therapy are not easily accessible. There are challenges in Ethiopia that prevent patients from

receiving CD4 testing, such as lack of infrastructure or laboratories at health facilities to support conventional testing, poor service in patients record management and data processing, reagents requiring cold chain transportation and/or storage, and weak networks between health centers and hospitals [6]. On the other hand, 24 percent of health facilities providing ART had conventional CD4 diagnostic machines by mid-2011 and only these facilities can provide same day results; even with specimen referral networks, this limits patient access to CD4 testing. Conventional CD4 testing through specimen referral requires patients return to the facility to receive their results [6].

1.2. Statement of the problem

Patients receive antiretroviral treatment (ART) which they need to take every day for the rest of their life [8]. To maintain treatment efficiency, it is necessary to start the treatment at a suitable time. Although the debate regarding when to start antiretroviral therapy has been going on for over two decades, consensus on this question has been hard to achieve. This lack of clarity continues in the current era, with major guidelines recommending very different treatment strategies [8]. A recent study using validated computer simulation to weigh important harms from earlier initiation of ART (toxicity, side effects, and resistance accumulation) against important benefits (decreased HIV-related mortality) found that earlier initiation of ART is often favored, although the TLC predicts survival in untreated patients, the exact cut-off point for beginning treatment remains controversial [9].

In resource-limited settings the decision to initiate ART in adults and adolescents relies on clinical and immunological assessment. In order to facilitate the rapid scale-up of ART programme with a view to achieving universal access to this therapy, WHO emphasizes the importance of using clinical parameters in deciding when to initiate it [5]. However, it is recognized that the value of clinical staging in deciding when to initiate and monitor ART is improved by additional information on baseline and subsequent CD4 cell counts [5].

In managing patients' record, although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. On the other hand, some of the challenges that affect the optimum time to start ART are motivation of patient due to accessibility and cost, inefficiency in patients' data processing and decision making, shortage of health counselors which create delay for HIV patients to start ART on time [10].

On the other hand since patients must return to the facility multiple times to receive CD4 results, there will be high burden on patients, and it can adversely affect effective management of patients on treatment and delay in initiation of treatment and cause patient loss-to-follow-up [6].

The decision as to when to start anti-retroviral therapy is made even more difficult when CD4 cell counts are not available or not easily accessible for patients [1]. Patients having relatively higher CD4 at the first diagnosis, with the resource limitation of the country might take long time to recheck again which makes many patients to start ART too late to gain optimal benefit from the treatment. Therefore, for pre-ART patient appointing for the next recheck for CD4 count having perfection is required from the physician to determine the accurate range of time. For this there must be a data mining tool that supports the physician based on the patient's information [11]. Practically a pre-ART patient can be on ART before a month, within 1-2 months or after two months. According to Kendon,et al three classification of ART starting time can be used. Patients who initiated ART before 28 days are considered as immediate, those who initiated between 29 and 56 days as early class and those who initiated after 57 days as delayed class. But still there is a question on which patient with a given attributes (information) is on immediate, early or delayed class.

There is no single proven model for delivering ART. In resource-poor settings, little is known about the effect of treatment on patient survival and quality of life [12]. Treatment guidelines are from the developed world. This highlights the urgent need for generating regionally suitable ART data processing model [12].

Taking the above challenges into consideration, this research is principally intended to develop a model which can predict when the optimal time to initiate ART for pre-ART patients. This will save time and resource especially for resource-limited countries like Ethiopia by adding value to the decision made by the physicians in following up the pre –ART patients. Accordingly, the research has attempted to answer the following research questions:

- Is it possible to use data mining technique to build predictive model for determining ART initiation time?
- Which attributes are more important to predict ART initiation time of pre-ART patients?
- What are the interesting rules to mine association between the selected attributes to determine the optimal ART initiation time?

1.3. General Objective

The general objective of this research was to apply data mining techniques to construct predict model that can be used for determining the initiation time for Antiretroviral Treatment (ART).

1.3.1. Specific Objectives

For the accomplishment of the stated general objective the research work has considered the following specific objectives;

- To generate quality dataset by applying different data mining pre-processing techniques.
- To identify the attributes which have strong association with determination of optimal time for ART initiation
- To mine different association rules that show relationship between the attributes that have strong relationship with ART initiation time.
- To build classification models for predicting ART initiation time for pre-ART patients
- To evaluate the performance of models constructed using test.

1.4. Scope and Limitations of the Study

This research scope limited to developing a predictive models for the Initiation Time of ART patients using Addama and Ambo hospital ART datasets. Moreover, the study is limited to build a prototype for ART Initiation Time prediction and use classification data mining technique. The classification algorithms selected by the researcher in this study to develop prediction models are Decision Tree, Naïve Bayes and PART. Besides the promising findings observed in this study, the following major limitations are recorded:

- This study was conducted in two government hospitals namely, Addama and Ambo hospitals due two shortages of time and financial constraint.

- The quality of data is so poor and documentation explaining about the dataset was so minimal.
- Another limitation of this study was lack of literatures related to the application of data mining techniques to prediction of ART Initiation Time of HIV patients.

1.5. Methodology

1.5.1. Research Design

To achieve the general as well as the specific objectives of the study, the hybrid of two data mining model has been used: Knowledge Discovery Process (KDD) and Cross-Industry Standard process (CRISP DM). This hybrid model has been utilized because it is capable of providing: more general, research-oriented description of steps in knowledge discovery process.
It does emphasize the Iterative features of the process, drawing experience from previous models. It also supports academia and Industrial data mining projects.

In this case, this study applied only the Apriori Algorithm for Association rule mining in order to optimize the space required for searching itemset while generating the intended rule.

1.5.2. Business/Problem Understanding

On this stage for the purpose of identifying the appropriate tools of Data Mining, in order to clearly specify the benefits of the research after defining the problems and the alternative solutions, review of related literatures and consulting domain experts has been done.

1.5.3. Data understanding and data collection

Data used to address the objective of this study has been collected from Adama and Ambo Hospitals from database containing list of patients who are on ART treatment. To understand the nature of the data and to solve data quality problems, discussion has been conducted with domain experts and other personnel who interact with the data. There were a total of 18216 patients with 81 attributes on the data that has been used for this

investigation and out of them identifying the important attribute that have relation with initiation time for starting ART treatment has been selected and used for this study.

1.5.4. Data preparation and Preprocessing

Preparation and preprocessing of data involves the process of identifying and handling general data quality issues which includes confirming for completeness, redundancy, noisy data value, missing value and outliers from the dataset. After feature selection and extraction algorithms process to acquire cleaned data has been undertaken a qualified data that can fulfill the specific input requirements for the Data Mining tool has been achieved.

1.5.5. Developing, Analyzing and Evaluating Data Mining Model

Development of the model has been done after going through different available modeling techniques and selecting the one which fits the data that were used for this research. Accordingly the developed model has been analyzed and evaluated for its efficacy based on different criteria's.

1.6. Significance of the study

The findings of this research could be used to predict the ART Initiation Time of HIV patients attending Pre-ART groups to start therapy by taking the ten predicting variables suggested by this study. This has a great deal of benefit to patients, physicians, hospitals, policy makers, researchers and local and international partners working on supporting the implementation of ART programs. Accordingly, the benefits offered to these parties are:

- Physicians can make use of the model to forecast ART Initiation Time of their patients ahead which in turn eliminates complications which might occur due to not knowing the CD4 count during the right time. In addition to this, in facilities with no CD4 counting machine it can be used as a replacement.
- Patients receive the right regimen associated to their CD4 level so that their immunity level develops to a level where no opportunistic infection can occur.
- Policy makers can make use of the model to develop new guidelines or modify the existing one in order to improve implementation of ART programs in the country.

- The result obtained in this study can be taken as a base to conduct clinical investigation to validate the findings with the real situations and also similar researches can be done in some other part of the country to validate the findings and then make it to serve at a national level.
- Local and international partners can make use of the findings to identify the kind of support expected from them.

1.5.6. Ethical Consideration

The study has been carried out after getting permission from the ethical clearance committee of Addis Ababa University, School of Public Health and School of Information. Written permission has been collected from Adama and Ambo Hospitals after Objective and purposes of the study has been discussed.

1.5.7. Organization of the Thesis

The organization of this thesis is as follows: Chapter 1 explains backgrounds of HIV/AIDS impact in public health, its medication and WHO Guide lines, General over view of ART, application of Data mining in ART, Current practice of ART in Ethiopia, statement of the problem, the anticipated outcomes, objectives of the study, and research methodology,

Chapter 2 illustrates Data mining over view, data mining and knowledge discovery in databases, data mining methodology used in this study, data mining process models (focusing on the six steps hybrid KDP data mining process model used as a framework for the current study), data mining techniques, algorithms and data mining and its applications to the Health care.

Chapter 3 explains specific activities performed to understand the data and preprocessing the dataset. Activities performed in the study include description of data source, selection of ART attributes, exploratory Data Analysis of the selected attributes, Data and Dimensionality Reduction Data Preparation (For Weka) binning/ Discretization, managing missing values, errors and resolving inconsistencies.

Chapter 4 describes all the experiments carried out to discover association rules and how to develop predictive models were explained. Multiple experiments were done by modifying the

parameters of each algorithm. All experiments have passed two stages of model development i.e. training and then validation of the models. Accordingly, results of the experiments are analyzed and interpreted.

Chapter 5 provides conclusions and recommendation of the study.

CHAPTER TWO
LITERATURE REVIEW

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a “data rich but information poor” situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools [13]. As a result, data collected in large databases become ”data tombs” data archives that are seldom revisited. Consequently, important decisions are often made based not on the information-rich data stored in databases but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data [13]. In addition, consider current expert system technologies, which typically rely on users or domain experts to manually input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly. Data mining tools which perform data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research [13]. The widening gap between data and information calls for a systematic development of data mining tools which will turn data tombs into “golden nuggets" of knowledge [13].

2.1 Data Mining Over View

Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis [14].

As Berry and Linoff [15] stated, data mining usually makes sense when there is huge amount of data. On account of this reason most of the algorithms developed for data mining purpose requires large volume of data so as to build and train models that are intended to be used for different tasks of data mining such as classification, clustering, and association rule discovery. The rationale behind the need for bulky data is simple and straightforward, small training data results in unreliable generalizations based on chance

patterns. As a result, most data mining tools and algorithms demand large amount of training data (data used for building a model) in order to generate unbiased models [15].

In healthcare, “data mining is becoming increasingly popular, if not increasingly essential” [16]. Several factors have motivated the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders [17]. Fraud detection using data mining applications is prevalent in the commercial world, in the detection of fraudulent credit card transactions. Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection [18]. Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods [17]. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care [16]. Benko and Wilson [19] also argue that healthcare organizations that perform data mining are better positioned to meet their long-term needs.

2.2. Data Mining and Knowledge Discovery in a Database

Data mining techniques can highly be important in the health sector as this sector generates huge and complex volume of data which makes un-automated analysis impractical and expensive [20].

In real world of healthcare exist insurance fraud, abuse and other related problems and data mining techniques can be applied to detect such behaviors by generating information which can be important for all stakeholders [20]. Data Mining techniques can serve the health sector in a number of applications such as evaluating treatment effectiveness, health care management, the analysis of relationships between patients and providers of care, pharmacovigilance, fraud and abuse detection [20].

The aim of data mining is to discover hidden and interesting patterns in data. Those patterns may not be easily recognized using traditional method [21]. Data mining techniques

can also be applied in the health sector to bolster the body of knowledge that might be helpful in identifying possible healthcare risks so as to undertake measures to mitigate probability of the problems before they happen [22].

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD) [22].

With the emphasis on collected data increasing around the world, there is an urgent need for a new generation of different techniques, methods and algorithms to assist researchers, analysts, decision makers and managers in extracting useful patterns from the rapidly growing volumes of data [22]. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD). KDD has evolved from interaction and cooperation among such different fields as machine learning, pattern recognition, database, statistics, Artificial Intelligence, knowledge representation, and knowledge acquisition for intelligent systems. The main idea in KDD is to discover a high level knowledge (abstract knowledge) from lower levels of relatively raw data, or to discover a higher level of interpretation and abstraction than those previously known [22].

The knowledge discovery process (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [23]. The process generalizes to non-database sources of data, although it emphasizes databases as a primary source of data. It consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain [23].

2.3 Data Mining Methodology

The ultimate goal of the KD Process (henceforth KDP) model is to achieve overall integration of the entire process with industrial standards. Another important objective is to provide interoperability and compatibility between the different software systems and platforms used throughout the process. Integrated and interoperable models would serve the end user in automating, or more realistically semi-automating, work with KD systems. The efforts to establish a KDP model were initiated in academia [24].

Although, the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa. The researcher restricts our discussion to those models that have been popularized in the literature and have been used in real KD projects which is proper to the researcher's business domain.

2.3.1. Hybrid Model

As mentioned earlier under chapter one the researcher has employed a hybrid Data mining Model that is, Knowledge Discovery Process (KDP) and Cross-Industry Standard process (CRISP DM) to achieve the general as well as the specific objectives of the entire research. Hybrid Data mining model provides more general, Research-oriented description of steps in knowledge discovery process. It also does emphasize the iterative features of the process, drawing experience from previous models and it supports academia and Industrial data mining projects.

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model developed by Cioset.et-a.[24] (see Figure 2.1). It was developed based on the CRISP-DM model by adopting it to academic research [24- 26].

A hybrid of the above-mentioned approaches may be considered in determining a suitable goal for DM. All KDD model process models emphasise the iterative nature of the process that a DM

application is conducted. Typically, goals are selected, an experiment is conducted, based on results at each stage, a step is revisited or moves to the next step [26].

The iterative nature of KDD model process models allow retracting and considering different approaches/paths (goals, techniques and methods) in conducting a DM experiment as a way to address this uncertainty. This approach, however, is not optimal and results in a trial-and-error process, which is resource-intensive and risky with no guarantee of favourable results. Approaches to minimise unsuccessful attempts and provide certain guarantees would be highly beneficial [26, 27].

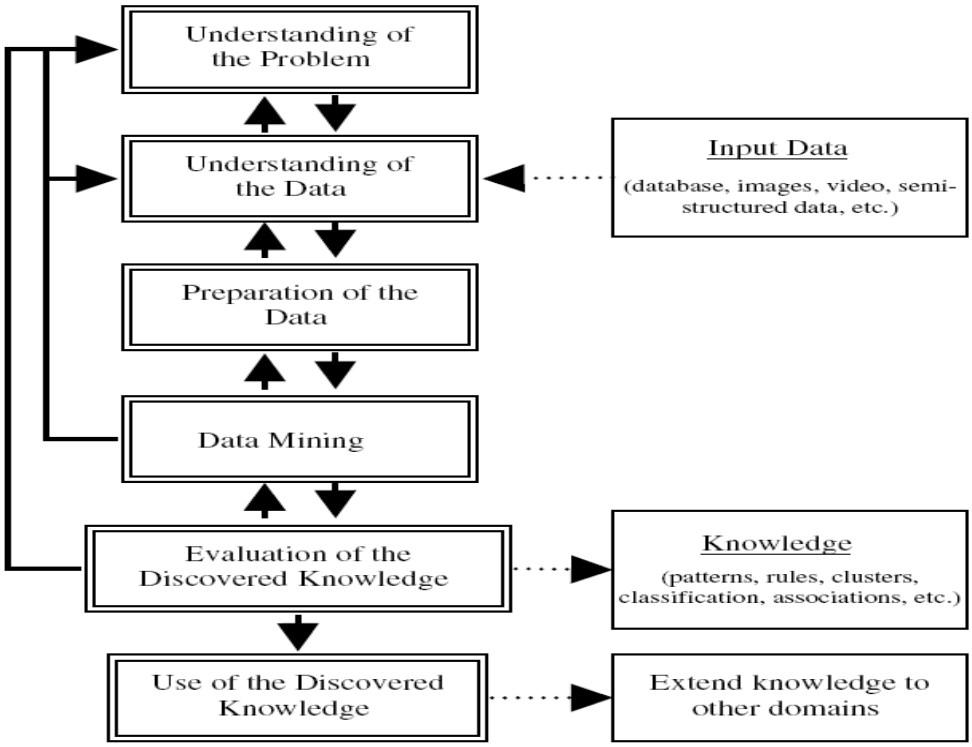


Figure 2.1: Hybrid Process Model

A description of tasks at each step in the six steps of hybrid KDP model are given as follows:

- **Understanding of the problem domain:** This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem.
- **Understanding of the data:** This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts.

- **Preparation of the data:** This step concerns deciding which data will be used as input for DM methods in the subsequent step.
- **Data mining:** Here the data miner uses various DM methods to derive knowledge from preprocessed data.
- **Evaluation of the discovered knowledge:** Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.
- **Use of the discovered knowledge:** This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains.

2.4. Data mining technique

2.4.1 Descriptive Model

Description and visualization can contribute greatly towards understanding a data set, specially a large one, and detecting hidden patterns in data, especially complicated data containing complex and non-linear interactions [15].

Descriptive model enable us to find natural groupings of the data and patterns that are interpretable and understandable by human beings. In other words, one can easily note that a descriptive model presents the main features of the data. It is essentially a summary of the data, permitting us to study the most important aspects of the data without their being masked by the sheer size of the data set. It is finding human-interpretable patterns, associations or correlations describing the data [22,27].

Han and Kamber [13] also used another term “sequential pattern mining” to refer to this type of frequent pattern mining in which searches are made to discover for frequent subsequences in a sequence dataset where attribute values show ordering of events. However, both clustering and sequential pattern mining are among the many variants of descriptive methods that can be applied when they are found fit for data mining objectives.

2.4.1.1. Association Rule Discovery

Association rules are a set of techniques, as the name implies, that aim to find association relationships in the data [28]. Association rule mining falls under the descriptive category. Association rules aim in extracting important correlation among the data items in the databases. Zhang et al. [29]. Have given association mining methods and the importance of rule interestingness measures. Association rule, basically extracts the patterns from the database based on the two measures such as minimum support and minimum confidence. To select the best measures for mining rules based on constraints such as multiple criteria is discussed in [30]. The support and confidence measures are stated in Jiawei and Micheline [31]. For mining frequent item set mining and association rule generation.

Support:

The rule $^{(A - B)}$ holds in the transaction set D with support s, where s is the percentage of transactions in D containing A U D

Support (A-B) =P (AUB).

Confidence:

The rule (A- B) has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contains B.

Confidence (A-B) =P (B\A)

In general, association rule mining can be viewed as a two-step process. The first step is generating all item sets having support factor greater than or equal to, the user specified minimum support. This is followed by generating all rules having the confidence factor greater than or equal to the user specified minimum confidence.

2.4.1.2. Association Rule Mining Algorithm

Association rule mining is performed using different algorithms of which The Naive Algorithm and Apriori Algorithm are most common.

Association algorithms used for the extraction of association rules from a set of transactions in TID-itemset format mine frequent itemset either after generating candidate itemset or without generating candidate itemset. Algorithms which generate candidate itemset for mining association rules usually have two successive phases in order to result into association rules. First, they find the frequent item sets. The aim of generating frequent itemsets is to extract all

sets of items from the transaction whose percentage of occurrence is greater than a certain minimum support S_{min} value. Since the data may consist of millions of transactions, and the algorithm may have to count huge number of potentially frequent (candidate) item sets to identify the frequent ones, this phase will be computationally expensive and challenging. Next, strong rules can be generated directly from the frequent item sets by taking those items whose confidence is greater than a minimum threshold value [24, 44, and 56]. Confidence as a measure of strength of a rule is the percentage of transactions in which the consequent is true when the antecedent is true.

The Apriori algorithm

The Apriori algorithm uses prior knowledge about an important property of frequent itemsets—hence its name. The Apriori algorithm takes advantage of the Apriori property to shrink the search space. The Apriori property states that if an itemset T is not frequent, then adding another item A to the itemset T will not make T more frequent. That is, if T is not frequent, T U A will not be frequent [56]. If a given itemset is not frequent i.e. if it does not satisfy the minimum support threshold, any superset of this itemset will also be infrequent [24]. The Apriori property is an antimonotone property i.e. if a set cannot satisfy a property, all of its supersets will also fail the same test. This helpful property is used to reduce the number of itemsets (the search space) that must be searched in every subsequent search to find frequent itemsets. The Apriori algorithm performs repeated search for frequent itemsets through the candidate itemsets, starting with 1-itemsets, through 2-itemsets, 3-itemsets and etc. [24, 44]. According to Cios et al. [24], the Apriori algorithm will follow the steps listed below in order to generate frequent itemsets.

– First finds all 1-itemsets
– Next, finds among them a set of frequent 1-itemsets, L1
– Next extends L1 to generate 2-itemsets (C_1: candidate itemsets each with 2 items)
– Next finds among these 2-itemsets a set of frequent 2-itemsets, L2
– and repeats the process to obtain L3, L4, etc.

2.4.2. Predictive Methods

Predictive data mining is the process of automatically creating a classification model from a set of examples, called the training set, which belongs to a set of classes. Once a model is created, it can be used to automatically predict the class of other unclassified examples [33].Predictive mining tasks, on the other hand, perform inference on the current data in order to make predictions [31].Moreover, Predictive techniques or methods focus on building a model that will permit the value of one attribute to be predicted from the known values of other attributes. It was observed that these methods could make use of two types of techniques on the bases of the type of values the designated attribute will assume. The first of these techniques used in predictive methods is classification which is appropriate when designated attribute is categorical. Numerical prediction (often called regression) is another method in which a model is built to predict a numeric value [34].

2.4.2.1 Classification

As stated by Han & Kamber [35], classification is a two-step process: learning and classification. During learning, a classifier is built describing a predetermined set of classes. The classification algorithm builds the classifier by analyzing a training set and their associated class labels. The learned model or classifier is represented in the form of classification rules. The accuracy of the classification rules is estimated using test data. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

Classification and regression are two data analyzing methods which determine important data classes or may construct models which can predict future data trends. The classification task predicts categorical values. For instance, while the classification model is constructed to categorize whether the bank loan applications are safe or risky, the regression model may be constructed to predict the spending of clients buying computer products whose income and occupation are given [36, 37].

In the classification and regression models the following techniques are mainly used: Decision Trees, Artificial Neural Networks, Genetics Algorithm, 4-K-Nearest Neighbor, Memory Based Reasoning and Navie Bayes are the mainly used techniques of in classification and regression

models [38]. In this study only two classification techniques selected by the researcher based on different benefits and advantages as stated here under:

2.4.2.1.1. Decision Trees technique

Decision trees provide a graphical representation of a tree with conditions associated to nodes that permit to classify a new instance in a predefined set of classes [39].

Decision tree are like those used in decision analysis where each non-terminal node represents a test or decision on the data item considered. Depending on the outcome of the test, one chooses a certain branch. To classify a particular data item, one would start at the root node and follow the assertions down until a terminal node (or leaf) is reached; at that point, a decision is made. Decision tree can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules [40].

Decision tree consists of nodes and branches connecting the nodes. The nodes located at the bottom of the tree are called leaves and indicate classes (see Figure 2.2). The top node in the tree, called the root, contains all training examples that are to be divided into classes. All nodes except the leaves are called decision nodes, since they specify decision to be performed at this node based on a single feature. Each decision node has a number of children nodes, equal to the number of values that a given feature assumes. All decision tree algorithms are based on Hunt's fundamental algorithm of concept learning. This algorithm embodies a method used by humans when learning simple concepts, namely, finding key distinguishing features between two Categories, represented by positive and negative (training) examples. Hunt's algorithms are based on a divide and conquer strategy. The task is to divide the set S, consisting of n examples belonging to c classes, into disjoint subsets that create a partition of the data into subsets containing examples from one class only. The following pseudo code summarizes the algorithm [23].

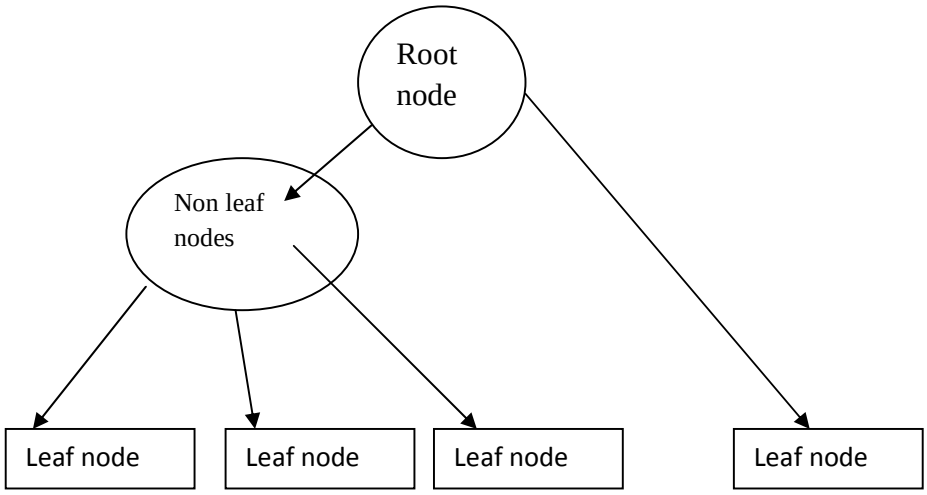


Figure 2.2: A Simple decision tree

2.4.2.1.2. Naïve Bayes algorithm

Naïve Bayes classifier: Provides an adaptive classifier that can improve initial knowledge-based predictions for the class of a new instance by refining the model on the basis of the evidences provided by the whole history of processed cases [39].

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class [34]. Naive Bayes is a type of supervised-learning module that contains examples of the input-target mapping the model tries to learn. Such models make predictions about new data based on the examination of previous data. The Naive Bayes algorithm uses the mathematics of Bayes' Theorem to make its predictions [34].

Bayes' Theorem is about conditional probabilities. It states that the probability of a particular Predicted event, given the evidence in this instance, is computed from three other numbers: the Probability of that prediction in similar situations in general, ignoring the specific evidence (this

is called the prior probability); times the probability of seeing the evidence we have here, given that the particular prediction is correct; divided by the sum, for each possible prediction(including the present one), of a similar product for that prediction (that is, the probability of that Prediction in general, times the probability of seeing the current evidence given that possible prediction) [34].

According to Han and Kamber [13], the naive Bayesian classifier works as follows:

- Let D be a training set of instances and their associated class labels. As usual, each instance is represented by an n-dimensional attribute vector, X = (x1, x2, … ,xn), depicting n measurements made on the instance from n attributes, respectively, A1, A2, … , An.
- Suppose that there are m classes, C1, C2, … , Cm. Given an instance, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier predicts those instances X belongs to the class C_i if and only if

$$P(Ci/x) = \frac{P(x/Ci)P(Ci)}{P(X)}$$

- As P(X) is constant for all classes, only P (X/Ci) P (Ci) needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P (C1) = P (C2) = … = P (Cm), and we would therefore maximize P (X/Ci). Otherwise, we maximize P (X/Ci) P (Ci). Note that the class prior probabilities may be estimated by P (Ci)=|Ci,D|/|D|, where |Ci,D| is the number of training instances of class Ci in D.

- Given datasets with many attributes, it would be extremely computationally expensive to compute P (X/Ci). In order to reduce computation in evaluating P(X/Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the instance (i.e., that there is no dependence relationships among the attributes). Thus,

$$P(X/Ci) = \prod_{k=1}^{n} p(Xk/Ci)$$

$$= P(x_1/C_i) * P(x_2/C_i) * \dots * P(x_n/C_i)$$

We can easily estimate the probabilities P(x1/Ci), P(x2/Ci), … , P(xn/Ci) from the training instances. Recall that here xk refers to the value of attribute Ak for instance X.

- In order to predict the class label of X, P (X/Ci) P (Ci) is evaluated for each class Ci. The classifier predicts that the class label of instance X is the class Ci if and only if

$$P(X/Ci)\ P(Ci) > P(X/Cj)\ P(Cj) \text{ for } 1 \leq j \leq m,\ j \neq i$$

In other words, the predicted class label is the class Ci for which P(X/Ci) P(Ci) is the maximum.

The Naive Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which can be used to calculate the probability of each of the possible classifications in turn. Having done this, the class with the largest value will be selected as the class of the new instance [41].

2.4.2.1.3. Rule Induction algorithm

As pointed by Larose Daniel [42] decision rule can be constructed from a decision tree simply by following a given path from the root node to any leaf. The complete set of decision rules generated from a class labeled dataset serve the same purpose as decision tree. Thus, decision rules are also called as classification rules [34]. Indicating that the rules can be used to predict the class of an unseen instance.

Rule induction algorithms generate a model as a set of rules logically ANDed together to form the rule antecedent (“IF” part) and the rule consequent (“THEN” part). The antecedent consists of the attribute values from the branches taken by particular path through the tree, while the consequent consists of the class value for the target attribute given by the particular leaf node [42].

PART algorithm combines the divide-and-conquer strategy (the top-down approach) for decision tree construction with the separate-and-conquer approach for rule learning. The separate-and-conquer strategy first builds a rule and then removes those instances that the rule covers. These

consecutive activities continue recursively for the remaining instances until none are left which generates sets of rules called ‘decision lists’ or ordered set of rules. On the other hand, in the partial decision tree, a pruned decision tree is built for part of the training instances, the leaf with the largest coverage is made into a rule, and the tree is discarded. Using partial decision trees in conjunction with the separate-and-conquer methodology adds flexibility and speed. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees. During the generation of such a tree, construction and pruning operations are integrated in order to find a “stable” sub tree that cannot be simplified further. Once this sub tree has been found, tree building ceases and a single rule is read off [2].

2.5. Model Evaluation Parameters

Evaluating the performance of a data mining technique is an essential feature of machine learning. Evaluation method is the benchmark to examine the efficiency and performance of any model. The evaluation is important for understanding the quality of the model or technique, foe refining parameters in the iterative processes of learning and for selecting the most acceptable model or techniques. Therefore, Data mining problems involving classification, it is very common to measure classifiers performance in terms of the error rate or misclassification rate [43].

The classifier predicts class label of each instances and if it is correct, it is counted as success, else counted as error. Evaluating the accuracy using training datasets derive a classifier or predictor to be likely misleading due to overspecialization of the learning algorithms to the data [43, 44]. For this reason it is better to assess the error rate based on independent test dataset that have no role in classifier datasets. Both, training data and the test data, needs to be representative sample of the problem [43]. For measuring accuracy of a classifier, there are a number of techniques such as the holdout, random sub-sampling, bootstrap and k-fold cross-validation, where the dataset is divided in to training and testing to train and test the classifier respectively [44].

As stated by Ian Witten and Eibe [43] the holdout method reserves a certain amount of instances for training and uses the remainder for testing (and sets part of that aside for validation, if required). In real-world terms, it is common to hold out one-third of the data for testing and use the remaining two-thirds for training. Bootstrap is based on the statistical procedure of sampling

with replacement [43]. The bootstrap procedure may be the best way of estimating the error rate for very small datasets.

In K-fold cross validation technique, one decide the number of fold (partitions of the data) and then the data is split in to K approximately equal partitions; and thus K-1/K and 1/K partition in turn used for training and testing the classifier respectively [43]. 10 fold cross validation is the most commonly used data partitioning technique for training and testing a classifier. Extensive tests on numerous different datasets, with different learning techniques, have shown that ten is the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up [43]. Although these arguments are by no means conclusive, and 42 debate continues to rage in machine learning and data mining circles about what is the best scheme for evaluation, tenfold cross-validation has become the standard method in practical terms [43]. Accordingly, ten-fold cross validation is selected for this research to train and test the classifier models.

2.5.1. Confusion Matrix

Confusion matrix is a tool for analyzing how well the classifier can recognize tuples of different classes [43]. Throughout this section the investigator had tacitly assumed that the goal of the performance evaluation was to maximize the success rate of the predictive model for ART dataset.

As stated by Vinterbo, S. A. [45] Predictive models are evaluated in terms of correctness, often referred to as performance, and applicability. The performance measures are almost always geared towards the evaluation of an instance of a model type, and are almost always realization method independent. Applicability measures also contain measures that apply to the model type itself, pertaining to the need of models to be evaluated in terms of their context [45].

Once a predictive model is developed using the ART dataset, the model should be checked as to how it will perform for the future data which, it has not seen during the model building process. The researcher used three different DM classifiers, techniques and tool to build the predictive model and in order to evaluate the performance of the model, confusion matrix and ROC analysis were used.

Moreover, the confusion matrix is a useful tool for analyzing how well the researcher's classifier can recognize tuples of different classes. The following procedures and rules were implemented to confirm the model performance evaluation for the results of the predicted model of ART Initiation Time. Given M classes; a confusion matrix is a table of at least size M by M. An entry, $CM_{i,j}$ in the first M rows and M columns indicates the number of tuples of class $_i$ that were labeled by the classifier as class $_j$. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry $CM_{1,\ 1}$ to entry $CM_{m,m}$, with the rest of the entries being close to zero [46].

In building a classification model, the confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models [47].

Table 2.1: Confusion Matrix

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

A confusion matrix table as shown above of size two by two, the following measures can be calculated to measure predicted pattern of the ART Initiation time model for ART dataset's accuracy of the model, True Positive Rate, False Positive Rate, Accuracy, Precision ,Recall, F-measure and ROC Curve.

The **True Positive Rate** of a classifier is expected by dividing the correctly classified positives by the total positive count.

$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$

The **True Negative Rate** of a classifier is estimated by dividing the incorrectly classified negatives by the total negatives count.

$$\text{True Negative Rate} = \frac{TN}{TN+FP}$$

The **Accuracy** of a classifier is projected by dividing the total correctly classified positives and negatives instances by the total number of samples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is calculated by dividing correctly classified instances by the total number of correctly and incorrectly classified samples.

$$\text{Precision} = \frac{TP}{TP+FP}$$

F-Measure is calculated as the harmonic mean of recall and precision.

$$\text{F-Measure} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

2.5.2. ROC Curve

Receiver Operating Characteristics abbreviated as ROC curves are a useful visual tool for comparing two classification models. a term used in signal detection to characterize the tradeoff between hit rate and false-alarm rate over a noisy channel [44].

ROC curves depict the performance of a classifier without regard to class distribution or error costs [44]. They plot the true positive rate on the vertical axis against the false positive rate on the horizontal axis [44].. The former is the number of positives included in the sample, expressed as a percentage of the total number of positives (**TP Rate = 100 × TP/ (TP + FN)**); the latter is the number of false positives included in the sample, expressed as a percentage of the total number of negatives (**FP Rate = 100 × FP/ (FP + TN)**). A sample ROC curve representing the percentage of true positives and false positives is presented in the figure below. The plot also shows a diagonal line where for every true positive of such a model, there is more likely to encounter a false positive. Thus, the closer the ROC curve of a model is to the diagonal line, the less accurate the model. If the model is really good, initially we are more likely to encounter true positives as we move down the ranked list. Thus, the curve would move steeply up from zero. Later, as we start to encounter fewer and fewer true positives, and more and more false positives,

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

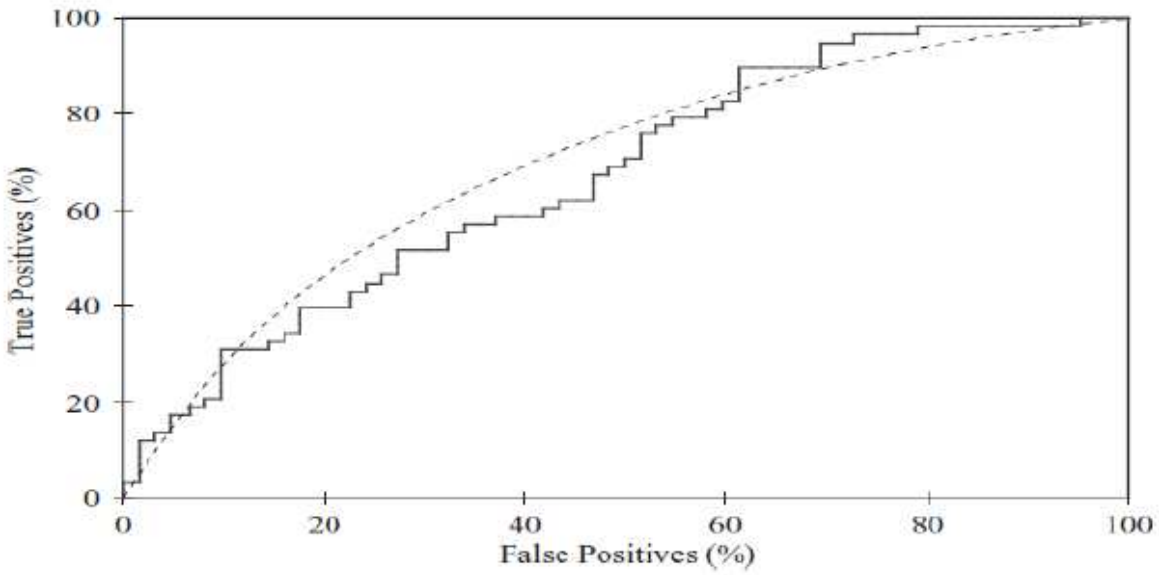


Figure 2.3: A Sample ROC Curve

2.6. Related works

2.6.1. Data mining application in Health Care

Medical data mining has been applied for accurate classification and rapid prediction for prognosis and diagnosis of patients in a specialized medical area [48]. It has been also used for training unspecialized doctors to solve a specific diagnostic problem [49]. Among several algorithms for classification and prediction tasks, a decision tree is one of the most frequently used techniques in medical data mining area. While it is easy to find many cases to prove the decision tree to be useful in the business domain, the decision tree enables to predict prognoses and diagnoses in the domain of medicine, using tree-structured models or in the form of 'IF condition-based-on attribute-values THEN outcome-value' to identify useful features of importance.

2.6.2 Mining HIV/AIDS database

Different studies have been conducted to assess the applicability of data mining on HIV/AIDS with various objectives [50, 51].

In Thailand, Wipada Chanthaweethip and SumanataGuha [50], conduct a research on Temporal Data Mining and Visualization for Treatment Outcome Prediction in HIV Patients in order to assist physicians monitoring HIV patients. Temporal data mining in this study has been applied to current patients being treated with ARV at HIV-NAT. ANN calculates the importance of a variable in achieving predicted results [50]. Finally, the researchers have indicated that the aim of their investigation has been achieved as visualization can provide overview monitoring by predicting treatment outcome. They have concluded that "User can see their patients who follow up at the clinic and can detect the event earlier, even for patients who do not visit at the clinic. This allows physicians to perform actions to prevent an event that may affect treatment efficacy [50].

The relationships of social, economic, and health care workforce and other factors on HIV/AIDS prevalence rates among the 194 countries were investigated using data derived from WHO and NGO sources. CART is used for prediction of the levels of HIV/AIDS prevalence rates after merging the data from the various data sources into one file at the country level for analysis. The results of study of the countries show that physician density is the first key discriminator between high and low HIV/AIDS prevalence rates. Standard ordinary least squares regression is performed using some of the same variables used in the CART approach, with the intent to build the best explanatory model [51].

The final model of Madigan et al. [51] revealed that the main factors in understanding HIV/AIDS prevalence rates are physician density followed by female literacy rates and nursing density in the country. In addition they argue that to reduce the impact of HIV/AIDS on health care systems globally require a multi-faceted approach. It may not be sufficient to simply request more financial support without revising current approaches to train and retain nurses and physicians and to increase female literacy. Governments may want to consider programs specifically targeted to strengthen and alternatively distribute the nursing workforce given that their finding showed nurse density having an association with HIV/AIDS prevalence rates [51].

Data mining research that examined the application of data mining technology to identify determinant factors of HIV infection and to find their association was conducted on 2005 by Abraham [52]. Out of the initial 82 attributes and 18646 records, case taken Center for Disease Control and Prevention (CDC), he used 19 attributes for decision tree classifier using knowledge

SEEKER algorithm of Knowledge STUDIO to identify risk factors. To generate association rule using Apriori Algorithms of Weka tool, 9 attributes and 5267 clients' records are used.

The above researcher 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. This research has attempted to identify determinants of HIV status by analyzing HCT data pattern using different and recent data and more variables so as to support the scaling up of knowledge of HIV status. The techniques used to identify determinants of HIV status are classification and association rule mining taking secondary data from ZVCT and OSSA-BHVCT centers. To the knowledge of the researcher this research problem has not been attempted before. Hence, this research is expected to contribute a lot in the scaling up of HIV counseling and testing [52].

A total of 250,000 records from HIV/AIDS patients were used to study the application of data mining techniques with the purpose of utilizing the data mining results for the management of HIV/AIDS in Thailand. IBM's Intelligent Miner is used for clustering and association rule discover and clustering is used in order to identify characteristics of categories of people affected by the disease whereas association rule mining is to identify symptoms that may follow a set of existing ones. The study is assumed to be contributing for the identification of patterns that can be used to target resources and for the monitoring of the disease. The findings of Vararuk et al.[53] (2008) showed that clustering revealed groups of patients with common characteristics and errors in the data. Association rules identified associations that were not expected in the data and are different from traditional reporting mechanisms utilized by medical practitioners. It also allowed the identification of symptoms that co- exist together [53].

CRISP-DM methodology and clustering and classification data mining techniques were used by Biru(2009), and attempted to investigate 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 was balanced only 14793 records was considered for his experiment. [53].

2.6.3. Application of data mining in Antiretroviral Therapy

A research conducted by Teklu [55], applied classification and association rules using, J48 and Apriori algorithms respectively, on a total number of 18740 ART patients' datasets in a study that 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. The methodology employed to perform the research work is CRISP-DM. Finally, the investigator proved the applicability of data mining on ART by identifying those factors causing the continuation or termination of the service [55].

A research conducted by Teklu [55], applied classification and association rules using, J48 and Apriori algorithms respectively, on a total number of 18740 ART patients' datasets in a study that 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. The methodology employed to perform the research work is CRISP-DM. Finally the investigator proved the applicability of data mining on ART by identifying those factors causing the continuation or termination of the service [55].

As it can be seen from the ART related studies conducted so far, it is possible to understand that there is no finding in relation to predicting ART initiation optimal time. Therefore, this study focuses on generating the association rules and constructing ART Initiation Time predictive model that help physicians to decide which attribute of patients should be associate with ART Initiation Time which can help physicians to manage Pre-ART patients' follow up more quickly and accurately.

CHAPTER THREE
UNDERSTANDING AND PREPROCESSING ART DATASET

3.1. Data Understanding

3.1.1. Description of the ART Data source

In order to achieve the objective of this research, collecting representative subset of ART data is a prerequisite. Therefore, Data involved in this study were based on the electronic medical records of HIV patients who were in follow up treatment in the ART department of Adama and Ambo Hospital ART Data bases. In both cases, the Data base is in SQL server and has the same format and interfaces 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 both Hospitals'. Raw data of 18216 records and 81 attributes. 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 our purpose of mining. Moreover, there have been attributes which are not clear for what they stand for and which the person who gave us 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 we 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.

3.1.2. Selection of ART Attribute

As stated by Deshpande and Thakare [48], deciding on the data that will be used for the analysis is based on several criteria, including its relevance to the data mining goals, as well as quality and technical constraints such as limits on data volume or data types. Therefore, in this thesis the attribute are selected with the help of domain expert and extensive literature review. Because taking all the variables in the data base we have, feed them to the data mining tool and find those

which are the best predictors may be does not work very well. Due to the time it takes to build a model increases with the number of variables and blindly including extraneous columns can lead to incorrect models [49].

Generally, the ART dataset in both Hospitals obtained contains many attributes. To decide on the relevant attributes for this study, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts in order to simplify the task of modeling.

As described above, the following attributes were selected from the entire data base: Age, Sex, Marital Status, Education Level, Religion, Family Planning, Occupation, OA Weight, OACD4, OAWHO Stage and Initiation Time are the final selected attributes which were prepared and preprocessed as stated in the following section. As shown in table 3.1 below the description of selected attributes, data types they take, list of values or range of values of these attribute, are given together with statistical summaries of these attributes in data description and exploratory data analysis section. The main objective of analyzing statistical summaries of these selected attributes is to see the distribution of each value of attributes in the dataset to identify errors (noises) and discern whether there exist missing values or not before training and developing the models.

Table 3.1 Selected attributes with their description

s/no	Attributes Name	Meaning	Values	Data type
1	Sex	Sex of the patient	Female, Male	Nominal
2	Age	Age of the patient	Numeric Age Value	Numeric
3	Marital Status	Marital Status of the patient	✓ Never Married ✓ Married ✓ Living together ✓ Divorced ✓ Separated ✓ Widower ✓ Other	Nominal
4	Educational Level	Educational Level of the patient	✓ No Education ✓ Primary ✓ Secondary ✓ Tertiary ✓ Other	Nominal
5	Religion	Religion of the patient	✓ Orthodox ✓ Muslim	Nominal

			✓ Protestant ✓ Catholic ✓ Other	
6	Family Planning	Family Planning Status of the patient	✓ True ✓ False	Nominal
7	Occupation	Occupation of the patient	✓ Employed ✓ Self Employed ✓ Un Employed ✓ Student ✓ Other	Nominal
8	OA Weight	The weight of the patient on ART	Numeric values ranged as <24 ,25-49, 50-73,74-98, 99-122, 123-146, >147	Numeric
9	OACD4	The CD4 count of the patient currently on the ART	<200, 200-349,350-499,>=500	Numeric
10	OAWHO Stage	WHO stage at which the patient is on ART	Stage 1 Stage 2 Stage 3 Stage 4	Nominal
11	Initiation Time	Patients who initiated ART at different days range.	Numeric values ranging from <28 days (Immediate), 29-56 days(Early),>=57 days(Delay).	Nominal

3.1.3. Selection of ART Instances

In order to build a predictive model for ART Initiation time requires selection of instances as well. Thus, in addition to the removal of irrelevant attributes which were done based on the attributes irrelevance to the prediction of ART Initiation time, instances that have undergone all HIV patients that are IN ART follow up alone were selected from the database. Out of the 18216 follow up victims, 11615 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 [44], 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 **175** records without class information were removed from subsequent analyses. The remaining dataset were then have **11440** records whose outcomes are distributed in one of the three outcome categories. Thus, the statistical summaries of attributes also relevant to the data mining objectives are on those **11440** records.

3.1.4. Exploratory Data Analysis

Descriptive data summarization techniques can be used to identify the typical properties of your data and highlight which data values should be treated as noise or outliers [44].The exploratory data analysis is performed to detect bad data i.e. attributes with the missing values and wrong entries or noises and inconsistency in values of attributes. In addition, through this technique it enables to facilitate the next phases of data preparation.

In this study, the attributes description, data type, unit of measurement and list of values or range of values are depicted with the use of frequency tables for (3) Numeric and (8) categorical variables as follows. The frequency tables for the selected attributes show the original distribution of values of attributes in instances of the dataset before any preprocessing is done on the dataset.

Sex Attribute: The sex attributes describes the sex of the patient. It is a nominal valued attribute the values to this attribute are Female and Male. Table 3.2.shows the statistical summary of these values the attribute has assumed in the dataset.

Table 3.2: Statistical summary for the Sex Attribute in the Dataset

Sex		Frequency	Percent
Valid	Female	6593	57.6
	Male	4779	41.8
Missing values		68	0.6
Errors/noises		0	0
Total		11440	100

Age Attribute: is an attribute used to show the age of individual patients that registered for ART follow up treatment in the hospital ART care service data base. Its IQR is 12. The most frequent age value is 30 years. The number of missing value for the attribute is 1364.The missing values

can be replaced by mean age which is 35 years. The number of values which were found outliers is **226**. The researcher decision on the outlier to remove and replaced by the mean value (35 years).

Table 3.3: Statistical summary for the Age Attribute in the Dataset

Age Attribute: Numeric			
Attributes		Frequency	Percent
Valid		10685	93.4
Missing		1364	11.9
Mean		35.05	0.3
Median		33.00	0.28
Mode		30	0.26
Std. Deviation		14.839	0.12
Variance		220.204	1.92
Range		892	7.79
Minimum		18	0.15
Maximum		910	7.95
Percentiles	25	28.00	0.24
	50	33.00	0.28
	75	40.00	0.34

Box plot graph is also plotted to visually depict the outliers. The third quartile (Q3) is 40; the first quartile (Q1) is 28. The Inter Quartile Range (IQR) is 12 as presented in the above table. 1.5*IQR is 18 years. Therefore, 40+18 is 58 years is the upper limit for outliers so that age values beyond 58 years are treated as outliers. For the lower limit is Q1-(1.5*IQR) which is 28-18=10. So Age values bellow 10 years can be considered outliers in this dataset.

Marital Status Attribute: is an attribute used to show the Marital Status of individual patient that registered for ART follow up in the hospital ART care service data base. It is nominal valued attribute and includes values as Never Married, Married, Living together, Divorced, Widower and Other. Table 3.4 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 3.4: Statistical summary for the Marital Status attribute in the Dataset

Marital Status		Frequency	Percent
Valid	Never Married	1449	12.6
	Married	5532	48.4
	Living together	1023	8.9
	Divorce	1025	9.0
	Separate	1459	12.8
	Widower	6	0.1
	Other	903	7.9
Missing values		43	0.4
Errors/noises		0	0
Total		11440	100

Educational Level Attribute: The educational level attribute used to show the level of education the patient has acquired. It assumes one of the five nominal values such as No education, Primary, Secondary, Tertiary and Other.

Table 3.5: Statistical summary for the Educational Level Attribute in the Dataset

Educational Level		Frequency	Percent
Valid	No Education	3396	2.0
	Primary	3727	32.6
	Secondary	2147	18.8
	Tertiary	1115	9.7
	Other	822	7.2
Missing values		233	2.0
Errors/noises		0	0
Total		11440	100

Religion Attribute: It indicates the religion of the patient. The Valid values of this attribute are nominal as follows: Orthodox, Muslim, Protestant, Catholic and Other.

Table 3.6: Statistical summary for the Religion Attribute in the Dataset

Religion		Frequency	Percent
Valid	Orthodox	7241	63.3
	Muslim	1281	11.2
	Protestant	2263	19.8
	Catholic	27	0.2
	Other	402	3.5
Missing values		226	2.0
Errors/noises		0	0
Total		11440	100

Family Planning Attribute: This attribute indicates the Family Planning status of the patient. Like the other attributes it assumes valid nominal values such as True and False. Table 3.7 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 3.7: Statistical summary for the Family Planning Attribute in the Dataset

Family Planning		Frequency	Percent
Valid	True	3321	29
	False	7995	69.9
Missing values		124	1.1
Errors/noises		0	0
Total		11440	100

Occupation Attribute: This attribute mainly indicates the occupation of the patient. Like the other attributes it assumes valid nominal values such as Employed, Self-Employed, Un Employed, Student and Other. Table 3.8 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 3.8 Statistical summary for the Occupation Attribute in the Dataset

Occupation		Frequency	Percent
Valid	Employed	1913	16.7
	Self Employed	3899	34.1
	Un Employed	3905	34.1
	Student	238	2.1
	Other	1110	9.7
Missing values		375	3.3
Errors/noises		0	0
Total		11440	100

OA Weight Attribute: The OA weight attribute indicates the weight of the patient on the ART service registered data base. It is Numeric data type attribute and Its Inter Quartile Range (IQR) is 53 as shown in the table below 1.5*IQR is 79.5 years. Therefore, 53+79.5 is 132.5 is the upper limit for outliers i.e OAWeight value beyond 132.5 kg. are treated as outliers. For the lower limit is Q1-(1.5*IQR) which is 0-79.5 = -79.5. So OAWeight values bellow -79.5 years can be considered outliers in this dataset. Table 3.9 shows the statistically summary of the OAweight attribute.

Table 3.9: Statistical summary for the OA Weight Attribute in the Dataset

OAWeight Attribute: Numeric		
Valid		10149
Missing		1291
Mean		37.91
Median		46.50
Mode		0.0
Std. Deviation		26.66
Variance		821.43
Range		1260
Minimum		-450
Maximum		810
Percentiles	25	0..0
	50	46.50
	75	53.00

OACD4 Attribute: OACD4 is an attribute which refers to the CD4 count of the patient who are in the care. It has 932 distinct values. The most frequent value for OACD4 is 0 to mean the CD4 count is not done for those patients. The researcher considered these values as missing values based on the discussion with domain experts. The mean value is 155 as calculated non 0 (non valid value). The zero value decided to be replaced by the mean value. The attribute needs discretization as its distinct values are too much. We can see table 3.10 below for important statistical summary. The frequency table is too long since we have 932 distinct values.to detect records with outlier values, five number summary was done on the none zero values i.e. only for those patients CD4 count was made accordingly;Q1 is 0 and Q3 is 126.the number of records detected as outlier was **876**.

Table 3.10: Statistical summary for the OACD4 Attribute in the Dataset

OACD4Attribute: Numeric	
Valid	11440
Missing	0
Mean	155.11
Median	0.0
Mode	0
Std. Deviation	441.65
Variance	195062.83
Range	8000
Minimum	0

Maximum		8000
Percentiles	25	0
	50	0
	75	126
	IQR	126
	1.5*IQR	189
	Upper limit	315

The Box plot graph is also drawn for OACD4 attribute in effort to find the outlier values for the attribute. The values we have above Q3 + (1.5*IQR) can be consider as outlier values as many literature recommends. In this case 126+189 =315. Therefore, values for the attribute will be removed and replaced by the mean value (155).

OAWHO Stage Attribute: Is an attribute used to show the OAWHO Stage of individual patient that registered for ART follow up in the hospital ART care service data base. It is categorical valued attribute and includes values Stage I, Stage II, Stage III and Stage IV. Table 3.11 shows the statistical summary of these values the attribute has assumed in the dataset.

Table 3.11: Statistical summary for the OAWHO Stage Attribute in the Dataset

OAWHO Stage		Frequency	Percent
Valid	Stage I	1096	9.6
	Stage II	3198	28.0
	Stage III	3561	31.1
	Stage IV	3585	31.3
Missing values		0	0
Errors/noises		0	0
Total		11440	100

Initiation Time Attribute: Initiation Time attribute is a derived attribute from ART Registration date and ART Start date of the patient. It is calculated as (ART Registration date of a patient-ART Start date of a patient) and named as Duration in days of a patient and have values “Immediate ART”, “Early ART”, and “Delayed ART” so that more general information will be obtained related to the Initiation time of the patients.

Table 3.12: Statistical summary for the Initiation Time Attribute in the Dataset

Initiation Time		Frequency	Percent
Valid	< =28	8828	77.2
	29-56	1186	10.4
	>=57	1426	1.5
Missing values		0	0
Errors/noises		0	0
Total		11440	100

3.2. Data Preparation and Preprocessing

The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of available data [24].Today's real world databases are highly susceptible to noisy, missing and inconsistent data due to the various reasons such as attribute of interest may not always be available, relevant data may not be recorded due to misunderstanding of the subject under consideration, instrument used may be faulty, etc…

There are numbers of data preprocessing techniques; Data cleaning, data integration, data transformation and data reduction are the most commonly used techniques that help to improve the overall quality of the data.

Data cleaning helps to remove the noisy data and correct inconsistencies in the data. Noisy data can be smoothed by the binary method, clustering or regression methods. Some data inconsistencies can be corrected manually by consulting the domain experts. Data integration usually applied to data from multiple sources in order to match up entities having similar properties. The other task in the preprocessing step is the data transformation technique. It is used to transform the data into a form appropriate for analysis.

3.2.1. Data and Dimensionality Reduction

Here this section dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data [56]. Though, 81 attributes in the dataset might take too much processing time especially with the algorithm Apriori, which consumes exponential processing time. Moreover as the likelihood of missing items increases with the increase in dimension, scarcity was the other problem observed. The other problem is

that, explained earlier, there are attributes containing values which can be seen as redundant, which also cause a waste of space and time. Therefore, columns containing redundant values and columns with lots of missing values which cannot be filled by comparing against other data values were removed from the dataset.

The pre KDD step of KDD process implies understanding the business before starting the process of data mining. These too have made contribution in reducing the attributes. Therefore, in addition to removing columns which have lots of missing and corrupted values, we also reduced attributes that are not so clear for us and which couldn't also be explained by the personnel who gave us the data. As the interest of data mining techniques goes usually with categorical data rather than data items identifying an individual, all attributes such as ID, telephone number, etc. that would single out an individual were excluded from the dataset. In the other way, this can also be taken as a move to protect individual privacy.

In addition, the investigator has used data encoding to transform the data represented in characters to digits, so that the original data is represented with a reduced form without losing any detail of the original data. Therefore, the following attributes were left for the purpose of association rule mining and for patients following ART Treatment at Adama and Ambo Hospital. The data encoding done on the study variables is depicted in the table below.

Table 3.13 Data Encoding of Continuous Numeric Attributes

s/no	Attribute's Name	Old-Values	New –Value
1	Sex	{Female, Male}	{1,2}
2	Age	{18-24, 25-29, 30-34, 35-39, 40-44, 45-49, and above 50}	{1, 2, 3, 4, 5, 6, 7}
3	Marital Status	{Never Married , Married, Living together, Divorced, Separated, Widower and Other}	{1, 2, 3, 4, 5,6,7}
4	Educational Level	{No Education, Primary, Secondary, Tertiary and Other}	{1, 2, 3, 4,5}
5	Religion	{Orthodox, Muslim, Protestant, Catholic and Other}	{1, 2, 3, 4,5}
6	Family Planning	{True, False}Family Planning usage status of patient	{1,2}
7	Occupation	{Employed, Self Employed, Un Employed, Student and Other}	{1, 2, 3, 4,5}
8	OA Weight	Numeric values ranged as: <24 ,25-49, 50-73,74-98, 99-122, 123-146, >147}	{1, 2, 3, 4,5,6,7}
9	OACD4	{<200, 200-349,350-499,>=500}	{1, 2, 3, 4,}
10	OAWHO Stage	{Stage 1, Stage 2, Stage 3, Stage 4}	{1, 2, 3, 4,}
11	Initiation Time	{<=28, days (Immediate), 29-56, days (Early)>=57, days (Delayed) }Numeric values ranging from <=28 days(Immediate)	{1,2,3}

3.2.2 Data Cleaning

3.2.2.1. Managing Missing Values

After limiting the dataset attributes in the manner above, the next task was dealing with rows (records) that contain missing values. The term Missing values refers to one or more fields of an attribute which have no value in it. The existence of many such cases makes the dataset incomplete and building models of any type whether Missing values come in the process of knowledge discovery not by human mistakes and omissions of data but also when data for certain variables is hard, costly or even impractical obtain [22]. Descriptive or predictive with incomplete data makes the resulting model none representative of the reality [22]. We would end up with little records if we keep deleting rows containing missing values. As it was learnt from section 3.1.4.above Exploratory Data Analysis step, with the use of descriptive statistical summaries, all the attribute are having missing values ordinate under each attribute in the attributes selected for this study are replaced automatically by a feature called "Replace Missing Values" in weka. "Replace Missing Values" replace the mode of nominal valued attribute for missing values and the mean of continuous valued attribute for missing values. Replacing the

mode or the mean is preferred method to removing an instance only because of a single missing value in on particular cell [12].The following Table 3.14. depicts the attributes, percentage of missing values and the handling mechanism that “ReplaceMissingValues” implements.

Table 3.14:The percentage of missing values and their handling mechanism for the selected Attribute.

No	Attributes Name	Missing values (%)	Handling mechanism
1	Sex	1.0	Replaced by the most frequent value.
2	Age	8.0	Replaced by the most frequent value.
3	Marital Status	0	Has no missing value
4	Educational Level	2.0	Replaced by the most frequent value.
5	Religion	2.0	Replaced by the most frequent value.
6	Family Planning	1.0	Replaced by the most frequent value.
7	Occupation	3.0	Replaced by the most frequent value.
8	OA Weight	11.0	Replaced by the most frequent value.
9	OACD4	0.75	Replaced by the most frequent value.
10	OAWHO Stage	0.73	Replaced by the most frequent value.
11	Initiation Time	0	Because all (175) instances with missing class information are ignored.

3.2.2.2. Noisy Correction

3.2.2.1. 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 [12].

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 have no jobs. Therefore for the sake of consistency, we corrected them be under one category “Unemployed”. There are also other examples of expressing an occupation by different words or spellings (correct and erroneous). Therefore we had to choose a single term that can serve instead of them. According to Han and Kamber [12], knowledge about the properties of the data can be used in detecting discrepancies that may exist in databases. With the help of the knowledge of the

possible values that each attribute can take, the same measure had been taken for the other attribute values and summarized as follows in the table.

3.2.2.4.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 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, the five - number summary (based on quartiles), the inter quartile range and the standard deviation. Box plots can be plotted based on the five number summary and are a useful tool for identifying outliers Han and kamber [12].

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 [12] 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

will be considered outliers values. Based on the recommendations Age, OA Weight and OACD4 which are numeric in data type, have a kind of outlier us stated under section 3.1.4. Exploratory Data Analysis. The outlier limits for the Age, OA Weight and OACD4 attribute lies between.

3.2.2.4. Binning/Discretization

Since discretization is a process of dividing the range of continuous attributes into disjoint regions (intervals) which labels can then be used to replace actual data values. Both in machine learning (ML) and data mining (DM) the discretization techniques are mainly used as a data preprocessing step, however they aim at different goals. In ML such techniques are usually applied in a classification context where the goal is to maximize the predictive accuracy.

As association rules work for categorical data only, there was the need to discretize all values and binning of the data under the field ‘Age’, ‘OA Weight’ and ‘OACD4’ We used Weka for this process. We also changed all numeric values to nominal so as to make the dataset ready for processing with Weka.3.6.version.

3.2.2.5. Description of preprocessed and prepared Data for Weka Tool.

So far different corrective measures were taken on the remaining attributed. After finishing the data cleaning process, we saved the file into csv format and lodging it to weka. The final summary of the dataset constructed ready for experiments with the use of algorithms is depicted as follows in table 3.15.

Table 3.15: Summary of the selected dataset Categories

Categories	**Description**
Number of attribute	11
Number of instances	11440
Number of classes	3
Size of the Data	1.39MB

CHAPTER FOUR
EXPERIMENTATION, ANALYSIS AND EVALUATION OF DISCOVERED KNOWLEDGE

Experimentation, in this study, represents the data mining step in the six step hybrid KDP model where four data mining algorithms (including the association algorithm) are applied on the dataset to achieve the objective of extracting association rules from attribute values of ART data base and to build a model for predicting ART initiation time. Evaluation, on the other hand, is concerned with evaluating the result of each experiment with its own measuring criterion. This section of the study presents all the experiments together with objective measures and based on knowledge of the domain area, and expert evaluation.

In this study both Associations rule mining and predictive model building experiments conducted in two sessions. Association rule mining experiments are carried out with the use of Apriori algorithm, specifically changing its parameters such as minimum confidence and minimum support in order to discover the relationship of the selected attributes with ART Initiation Time. Likewise, experiments which make use of different classification algorithms are intended to build ART initiation time predictive model of relatively better sensitivity and specificity as compared to others.

4.1. Experimental Design

In this study, all experiments are done based on the final processed dataset which contains 11,440 instances and 11 attributes. In all experiments, eleven selected attributes (Sex, Age, Marital Status, Educational Level, Religion, FamilyPlanning, Occupation, OAWeight, OACD4, OAWHOStage, and InitiationTime) are used. Out of these attributes, the class or dependent attribute is the Initiation Time) for both associations rule mining and predictive model building the same processed dataset also used. The algorithms used during both predictive model building and association rule mining experimentations are found in Weka 3.6.version. This version works on many file formats than its antecedents and it is compatible with CSV file format. Thus, no additional effort was exerted to change the dataset from excel to ".arff" file format which is necessary in the previous versions. The prepared dataset is saved using CSV file extension format.

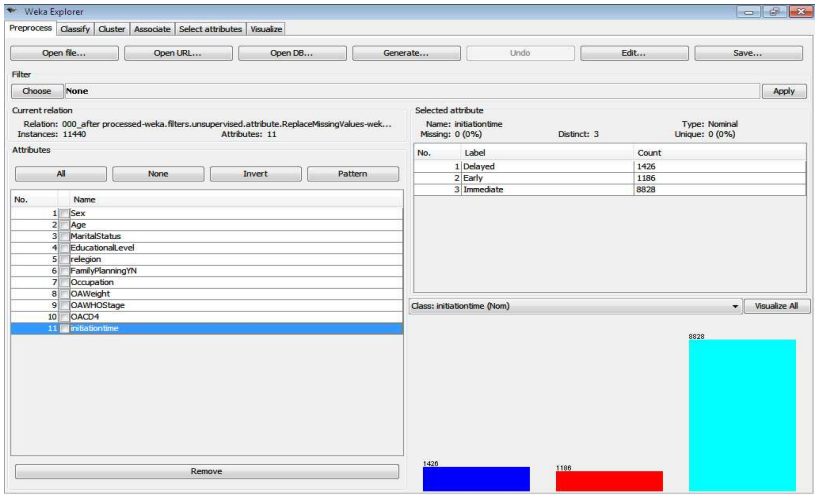


Figure 4.1: Weka Explorer window showing the number of attributes and instances.

1.1.1. Experimentation and Analysis of Association Rules

Association rules were the first data mining technique to be used for this research. After making the dataset with 11440 records ready for processing and loading it to Weka, we experimented association rule models were built at minimum support of 0.3 to 0.1 and with different minimum confidence varying from 50% to 100% thresholds. More specifically Table 4.1.shows the number of experiments to be carried out with their parameters.

Table 4.1.Experiments made for Association Rule Mining

Experiments #.	**Minimum Thresh holds**		**#rules generated (by default required=100)**
	Support	**Confidence**	
1.	0.3	50%	14
2.	0.25	60%	24
3.	0.25	70%	17
4.	0.25	80%	12
5.	0.2	80%	26
6.	0.2	85%	13
7.	0.2	90%	11

8.	0.15	85%	28
9.	0.15	90%	21
10.	0.15	95%	19
11.	0.1	90%	40
12.	0.1	95%	34
13.	0.1	100%	24
			283

So there will be 13 scenarios/Experiments at the end of the association rule mining using Apriori algorithm in weka. The number of rules generated for those experiments are also going to be analyzed, interesting rule will be obtained after sorting them with their confidence.

The association rule mining as it is well known produces many redundant rules. All the best rules found are sorted by their antecedents in order to identify redundant rules. Out of 283 best rules, 257 rules are either generated twice or more, or rules related with attributes which have insignificant strength with the dependent variable. Finally by removing the duplication and the unnecessary rules associated with insignificant attributes, 26 rules are taken. These rules are organized in the following for the purpose of analysis of the rules.

4.1.1.1.Association Rules Grouped by Education Level

.

Table 4.2: Association Rules by Education level

Rule#	Antecedents	Consequents	Confidence	Suppurt
1	EducationalLevel=No Education	initiationtime=Immediate	0.8	0.2
2	EducationalLevel=Primary	initiationtime=Immediate	0.8	0.2
3	EducationalLevel=Primary OAWHOStage=Stage IV	initiationtime=Immediate	0.95	0.1

As shown in the above table, Rule 1 and rule 2 depict those HIV patients who have no education and those having primary education level have the same probability of starting ART immediately with 80% probability covering 20% of the total HIV patients. On the other hand rule 3 shows that HIV patients with primary education level and in Stage IV of WHO stage has the highest

probability of starting ART immediately with 95% confidence covering 10% of the all HIV patients.

4.1.1.2.Association Rules Grouped by AOCD4

Table 4.3: Association Rules by AOCD4

Rule#	Antecedents	Consequents	Confidence	Support
1	OACD4='(-inf-31.5]'	initiationtime=Immediate	0.95	1
2	OACD4='(189-220.5)' Sex=F	initiationtime=Immediate	0.5	0.3

From the table above, Rule 1 depict that an HIV patient with CD4 count less than 31.5 has to start ART with high confidence (95%) covering 10% of the patients. On the other hand a Female HIV patient with CD4 between 189 and 220.5 has to start ART with 50% confidence and 30% support.

4.1.1.3. Association Rules Grouped by OAWeight

Table 4.4: Association Rules by OAWeight

Rule#	Antecedents	Consequents	Confidence	Support
1	OAWeight='(40.6-57.4]'	initiationtime=Immediate	0.5	0.3
2	OAWeight='(40.6-57.4]' OAWHOStage=Stage IV	initiationtime=Immediate	0.9	0.1

As indicated in the above table Rule 1 indicates that an HIV patient weight between 40.6 and 57.4 has to start ART immediately with 50% confidence and 30% support. But if an HIV patient with the same weight range is in Stage IV he/she has to start ART with 90% confidence covering 10% of HIV patients.

4.1.1.4. Association Rules Grouped by OAWHO Stage

.

Table 4.5: Association Rules by OAWHO Stage

Rule#	Antecedents	Consequents	Confidence	Support
1	OAWHOStage=Stage IV	initiationtime=Immediate	0.85	0.2
2	OAWHOStage=Stage IV '(189-220.5)'	initiationtime=Immediate	0.8	0.25

As shown in the above table, if a HIV patient is in Stage IV he/she has to start ART immediately with 85% confidence and 20% support. Similarly HIV patient in WHO stage IV and with CD4 between 189 and 220.5 has to start ART immediately with 80% confidence and 25% support.

4.1.1.5. Association Rules Grouped by Occupation

Table 4.6: Association Rules by Occupation

Rule#	Antecedents	Consequents	Confidence	Support
1	Occupation=Self Employeed	initiationtime=Immediate	0.7	0.25
2	Occupation=Self Employeed OAWHOStage=Stage IV	initiationtime=Immediate	0.95	0.1
3	Occupation=Unemployed	initiationtime=Immediate	0.7	0.25

As shown in above table, Rule 1 and rule 3 indicate that self employed and unemployed patients have to start ART with the same confidence (70%) and 25% support. But self-employed HIV patient who is in Stage IV more recommended to start ART immediately with 95% confidence and 10% support.

4.1.1.6.Association Rules Grouped by Sex

Table 4.7: Association Rules by Sex

Rule#	Antecedents	Consequents	Confidence	support
1	Sex=F	initiationtime=Immediate	0.8	0.2
2	Sex=F EducationalLevel=No Education	initiationtime=Immediate	0.9	0.1
3	Sex=F EducationalLevel=Primary	initiationtime=Immediate	0.85	0.15
4	Sex=F OAWeight='(23.8-40.6]'	initiationtime=Immediate	0.9	0.1
5	Sex=F OAWeight='(40.6-57.4]	initiationtime=Immediate	0.8	0.2
6	Sex=F OAWHOStage=Stage IV	initiationtime=Immediate	0.85	0.15
7	Sex=F Occupation=Self Employeed	initiationtime=Immediate	0.85	0.15
8	Sex=F Occupation=Unemployed	initiationtime=Immediate	0.85	0.15

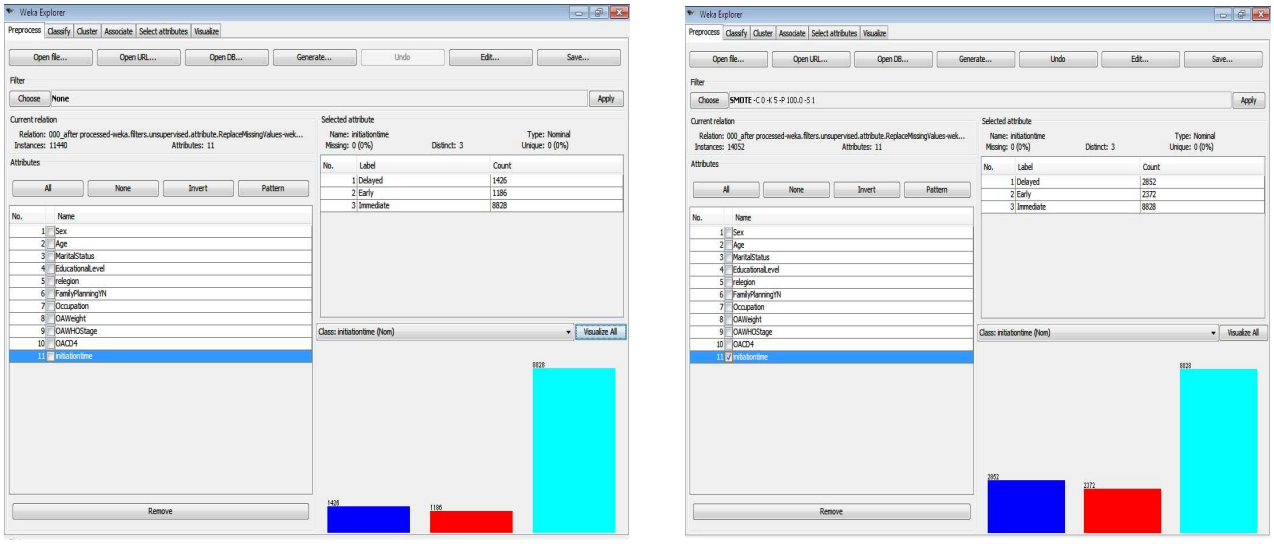
9	Sex=M	initiationtime=Immediate	0.6	0.25
10	Sex=M OAWHOStage=Stage IV	initiationtime=Immediate	0.9	0.1

As shown in the above table compared to male HIV patients, female HIV patients have to start ART immediately with better assurance which is 80% and 60% confidence respectively. Similarly female HIV patients with no education should take ART immediately with 90% confidence and 10% support compared to female HIV patient with primary education level (85% confidence and 15% support). Concerning the weight difference between female HIV patients, those with less weight are expected to take ART Immediately. That is Rule 4 and rule 5 depict Female HIV patient with weight in the range of 23.8 to 40.6 are expected to start at 90% confidence which is greater than the weight of female HIV patient with weight in the range of 40.6 to 57.4 with confidence 80%. From rule 6 and 10male HIV patients in stage IV are more expected to start ART immediately compared to female HIV patients in the same stage with confidence of 90% and 85% respectively. Finally rule 7 and 8 has depicted that both self employed and unemployed female HIV patients are equally expected to start ART immediately with 85% confidence and 15% support.

1.2. Experimentation for predictive Model Building

In case of developing a predictive model in datasets with high class imbalance and multiple classes requires some kind of countering the imbalance. Otherwise, simple comparison of models with accuracy alone may result in high predictive accuracy but low sensitivity and specificity In addition, Performance of DM algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large. In the case of ART dataset, the three class variable status has a higher imbalance. Therefore, the researcher used Synthetic Minority Oversampling Technique (henceforth SMOTE) automatic operation by filter where minority classes are over sampled 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 [34] As follows Figure 4.2 shows side by side the imbalance among the class attribute review of the class attribute status after SMOTE operation applied to the minority class.

Originally there were 8,808 records in the majority class (Immediate) and 1,186 and 1,426 only records in the minority class “Early” and “Delayed” respectively. But after applying SMOTE the difference between the three were reduced only to 8828,2372 and 1426 records.



(a) Original data ”Before “ (b) “After” Balanced data using SMOTE.

Figure 4 .2 shows the imbalance(Before) and Balanced ”After” among the classes attribute.

4.2.1. Experimentation with J48 Algorithm

The J48 decision tree C4.5 algorithm builds decision trees from a set of predefined training dataset using the concept of information entropy and attribute ordering. In addition to using the default parameter settings of the algorithm to build predictive model with J48, an attempt was made to find better classifier by varying its important parameters.

Two experiments are conducted using J48 by switching the parameter with pruning to TRUE and FALSE to form two separate experimental settings. The Meta classifier algorithm (AdaBoostM1) is used in both scenarios to evaluate the performance gained in those minority classes. This indicates that a total of four experiments were conducted in this category. The confidence factor in all scenarios is made to be 0.5, which is found to be a better value after attempting successive experiments at different confidence levels. It is also confirmed by different researchers for its better accuracy than taking the default confidence value 0.25 [58, 59].

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

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

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

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

In the first scenario of this experiment, the 10 attributes and **11,440 instances** are used by taking the default parameter value with pruning. The result showed that, the experiment has generated a model with a tree size of **328** and **269** leaves.

In the second scenario, the same number of attributes and records are used to run the experiment. But relatively larger tree having a size of **1,043** and **886** leaves is generated.

In the third and fourth experimental setups a classifier with relatively better predictive accuracy is generated due to the addition of the boosting algorithm. The tree and leaf size of the third and fourth scenarios increased to **733** and **593,** and **1121** and **953** respectively. The performance of the four experiments is shown in the table below.

Table 4.8: J48 Experiments Performance Evaluation for the ART Initiation Time

Experiment	**Accuracy**	**Tree Size**	**Leaf Size**	**WTP Rate**	**WFP Rate**	**WPrecision**	**WF-measure**	**WROC Area**
J48 pruned	**96.19**	**328**	**269**	**0.962**	**0.02**	0.963	0.962	0.989
J48 Unpruned	95.53	1043	886	0.955	0.029	0.956	0.955	0.975
AdaBoostM1with + pruned J48	95.49	733	593	0.955	0.033	0.955	0.955	**0.993**
AdaBoostM1with + Unpruned J48	**95.62**	**1121**	**953**	**0.956**	0.029	0.956	0.956	0.992

As showed in the table above, the Accuracy and WTPR all models indicate the performance of the model in accurately classifying new instances in classes of ART Initiation Time and it is calculated to be: 96.19% with misclassification of 3.81% and 0.962 for the first scenario; 95.53% of correct classification with misclassification of 4.47% and 0.955 for the second scenario; 95.49% of correct classification with error rate of 4.51% and 0.955 for the third scenario and the final scenario has exposed 95.62% correct classification and error rate of 4.38% and 0.956 WTP Rate accordingly..

The false positive rate indicated in each model shows the percentage of records which are wrongly classified in to any of the three classes. Accordingly, the first model has wrongly classified 2% of the records and hence it is the least in this category.

The ROC area also indicates the tradeoff or the area under the axis of true positive and false positive rates. Therefore, as the area under the ROC curve gets larger, it indicates that the classifier is putting more true positives than false positives in the given class. However the first and the fourth experiments has resulted in predictive accuracy of 96.19% with 0.989 WROC and 95.62% with 0.992 shows that this experiment has very low sensitivity and specificity. Greater sensitivity and specificity among these experiments is observed in experiment three (AdaBoostM1with + pruned J48) relatively low accuracy of 95.49 with 0.993 great WROC. AdaBoostM1with + pruned J48 decision tree has scored a better performance taking the performance parameters indicated above. Therefore, the AdaBoostM1with + pruned J48 decision tree has been selected to be compared with other classifiers generated under this category. In a table presented under *Appendix E*, the *TPR*, *F-measure* and *area under the ROC curve* of the minority classes (**Delayed** and **Early**) has dramatically increased whereas the FPR has lowered to a smaller value which indicates that less number of instances are wrongly classified under those minority and majority classes.

4.2.2. Experimentation with Naïve Bayes Algorithm

Bayesian methods are based on assumptions of probability. The naïve bayes algorithm assumes the attributes are independent. The probability of co-occurrence of an attribute value together with a particular outcome value is computed. Then, the class of a new instance will be computed by multiplying the probabilities of values the instance has assumed under each attribute. Section 2.4.1.1 of chapter two discusses the general procedure that the naïve Bayes algorithm follows in

identifying the probabilities of the attribute values together with how the probability of a class is computed in order to predict the class of a new instance.

The most important parameter in relation to this study is displayModelInOldFormat. However, there are also other parameters which can be adjusted according to needs of data used in different research areas. Table 4.9 shows the description of the parameter and type of values it takes. The default value to this parameter is "False". The researcher has changed this value to "True" as displaying the model in old format is recommended to output the classifier's result for multi-valued class classification.

Table 4.9: Naïve Bayes classifier Experimentation with modifying its parameter

Experiment	**Accuracy**	**WTP Rate**	**WFP Rate**	**WPrecision**	**WF-measure**	**WROC Area**
NaiveBayes	**92.96%**	**0.93**	**0.021**	0.932	0.93	0.991
NaiveBayes-o	92.96%	0.93	0.021	0.932	0.93	0.991

The above Table shows some of the performance measures for Naïve Bayes with default values for its Parameters.The first Experiment using Naïve Bayes and effect of altering the value of displayModelInOldFormat to "True" on the models performance of using NaiveBayes-o. Irrespective of the schemes applied, Naïve Bayes resulted in accuracy of 92.96%, and WROC area of 0.991 for both experiments.

4.2.3. Experimentation with PART Algorithm

PART algorithm extracts rules. Due to this reason the algorithm is categorized under classification by rule induction. The detailed procedure of this algorithm in extracting rules is explained in chapter two. The algorithm builds partial decision trees and reads a path from the root of the tree to the leaf to read of a rule. The rules are ANDed together to give a complete set of rules. PART has almost a similar set of parameters with J48 algorithm that can be adjusted to build better model from datasets. Like the J48 experiments done above, PART experiments are also done in four experimental settings based on the parameter pruning and the boosting algorithm. The experimental settings are indicated based on the next scenario.

Setting #1: PART Experiment with pruning.
Setting #2: PART Experiment without pruning.
Setting #3: AdaBoostM1 Experiment with its default parameters and taking PART with pruning as a base classifier.
Setting #4: AdaBoostM1 Experiment with its default parameters and taking PART without pruning as a base classifier.

In this setting, the value assumed to all the performance measure is partially difference observed among the model. So it makes comparison necessary to select the one with relatively better measures of performance. Performance measures such as WROC and the WTPR are better in the third experiment than the other algorithms. Therefore, the model from the third experiment i.e. AdaBoostM1with pruned PART has an accuracy of 95.62% which is a little bit lower than the first experiment and WROC of 0.994, WTPR of 0.956 which is better than all the others.

In the first setting of PART experiment a classifier with an accuracy of **95.73%**, weighted TPR of 0.957 and weighted FPR of 0.957 is generated by taking the default value of pruning. In addition to these performance parameters, the model has generated a total of **172 rules** to represent the patterns found within the dataset.

In the second setting, the same number of attributes and records are used by switching the default parameter value of unprunned to "TRUE". This experiment has produced a classifier with an accuracy of **95%** and also it has relatively lower, TPR and higher FPR than the first and third experiment.

In the third experiment, a boosting algorithm is applied on the base classifier with the pruning state turned on. Accordingly, a classifier with better accuracy 95.62% from the first and third experiment. This experiment is by far better than those three experiments having higher WROC Rate.

The last experiment is again a similar boosting experiment done by using PART rule induction algorithm. But here, the changed parameter is the pruning state, which is made to be "TRUE", i.e. pruning do not happen during model development. This experiment has performed 95% accuracy and also it has TPR and FPR 0.9 and 0.033 respectively.

Therefore, the model from the third experiment i.e. **AdaBoostM1with pruned PART** has an accuracy of 95.62%, and higher WROC of 0.994 which is better than the others three experiments.

Therefore, from the above four experiments we can understand that One AdaBoostM1with pruned PART experiments have good predictive WROC than those done using the classifier. The overall performance of each experiment is presented in the table below.

Table 4.10: PART Experiments Performance Evaluation for the ART Initiation Time

Experiment	Accuracy	No. of rules	WTP Rate	WFP Rate	WPrecision	WF-measure	WROC Area
PART Pruned	95.73%	172	0.957	0.026	0.957	0.957	0.972
PART unpruned	95%	462	0.95	0.033	0.95	0.95	0.972
AdaBoostM1with pruned PART	95.62	272	0.956	0.027	0.957	0.956	0.994
AdaBoostM1with unpruned PART	95%	462	0.95	0.033	0.95	0.95	0.972

4.2.4. Model Evaluation

In this research project work, several experiments had been carried out with three classification algorithms, i.e. J48 decision tree algorithm, Naïve Bayes classifier and the PART algorithm to build a predictive model that predicts the optimal time of ART Initiation time in ART Dataset. From the experiments all attributes were identified to make sound rule and better accuracy.

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

Table 4.11: The selected Models Comparison for the ART Initiation Time

No.	Model	Accuracy	Precision	F-Measure	Mean Absolute Error	TPR	FPR	AUC
1.	AdaBoostM1with + pruned J48	95.49%	**0.955**	**0.955**	**0.0306**	**0.955**	**0.033**	**0.993**
2.	NaiveBayes	92.96%	0.932	0.93	0.0664	0.93	0.021	0.991
3.	AdaBoostM1withpruned PART	**95.62%**	**0.957**	**0.956**	**0.0301**	**0.956**	**0.027**	**0.994**

Considering the above numerical values labeled under each of the evaluation parameters, the boosted and prunned PART model has revealed a better performance. Hence, accuracy only can't be a valid qualification criteria for imbalanced datasets, the mean absolute error which is the average error calculated on those tests during ten iterations is a very good criteria both from data mining and statistical perspective [25]. In addition to mean absolute error, AUC is also a recommended parameter to evaluate model performance in the case of such imbalanced dataset.

The ROC area is the area indicating the proportion of true positives versus false positives by putting the TPR on the Y-axis and FPR on the x-axis. The higher the TPR and the lower the FPR

indicates maximum ROC area which is again an indication of good classifier. Therefore, the prunned and boosted PART model has revealed better performance in the above two parameters (Mean Absolute Error and AUC).

Therefore, for the given data under study taking the mean absolute error, FPR and ROC area; the boosted and prunned PART has shown better performance and has been selected as a best classifier model for the ART Initiation Time. The rules generated by this model are also used for interpretation and the Weka output of the model is appended in *Appendix E*.

4.2.5. Rule generated from the selected Model

Boosted and prunned PART rule learner with the specified scheme has resulted total of 272 rules. Out of these the rules which are highly predictive are selected and discussed as the finding of this study based on relevant to the domain. The following are selected best rules generated from the identified model.

Rule #1: If Sex = Male and OACD4 = 157.5-189 and Occupation =Self Employed and Age= 34-38and OAWHO Stage = Stage III, then the class of ART Initiation Time of the patient is likely expected to be Immediate for ART (203.44/85.97).

- The level of assurance of the independent attribute for the status or the predicted class in the bracket can be calculated as follow:

=> 203.44/ (203.44+85.9)= 203.44/289.34= 0.70= 70%

Rule #2: If Sex = Male and Occupation = Self Employed and Age =34-38 and Family planning YN=False and Educational Level=Primary and OAWHO Stage = Stage III, then the class of ART Initiation Time of the patient is likely expected to be Immediate for ART (147.58/58.73).

Rule #3: If OAWHO Stage=Stage III and Marital Status=Married and OA Weight=40.6-57.4, then the class of ART Initiation Time of the patient is likely expected to be Immediate for ART (406.82/20.21).

Rule #4: If Educational Level=tertiary and Occupation Employed and Marital Status=Married OACD4 =126.157.5 and Age=26-30, then the class of ART Initiation Time of the patient is likely expected to be Immediate for ART (169.65/4.94).

Rule #5: If Marital Status=Married and Occupation = Self Employed and Age=26-30, then the class of ART Initiation Time of the patient is likely expected to be Immediate for ART (512.34/312.58).

Rule #6: If Sex = Female and OAWHO Stage = Stage IV and Family Planning YN=False and Marital Status=Married, then the class of ART Initiation Time of the patient is likely expected to be Immediate for ART (33.0/8.0).

Rule #7: If Sex=Male and Educational Level = Primary and Age 26-30 and OAWHO Stage= Stage III and OACD4 = 126-157.5 and OA Weight= 40.6-57.4 and Marital Status=Married, then the class of ART Initiation Time of the patient is likely expected to be Early for ART (3821.64/1862.72).

Rule # 8: If Sex=Male and OAWHO Stage = Stage II and OACD4 126-157.5 and Marital Status =Married and Age=34-38 and Educational Level=tertiary, then the class of ART Initiation Time of the patient is likely expected to be Early for ART (279.39/123.74).

Rule # 9: If Sex=Male and OACD4 189-220.5 and Age=34-38 and Marital Status =Married and Educational Level=No Education and Occupation=Self Employed, then the class of ART Initiation Time of the patient is likely expected to be Early for ART (182.14/73.05).

Rule # 10: If Sex=Female and OACD4 126-157.5 and OAWHO Stage=Stage IV and Educational Level=No Education and Occupation=Unemployed, then the class of ART Initiation Time of the patient is likely expected to be Early for ART (108.75/5.12).

Rule # 11: If Sex=Female and OAWHO Stage =II and OACD4= 189-220.5 then the class of ART Initiation Time of the patient is likely expected to be Early for ART (17.0/1.0).

Rule #12: If Sex = Male and OACD4 = 126-157.5 and OAWHO Stage = Stage I and Occupation = Employed, then the class of ART Initiation Time of the patient is likely expected to be Delayed for ART (71.0/2.0).

Rule #13: If Sex = Male and OACD4 = 126-157.5 and OAWHO Stage = Stage II and OA Weight = 23.8-40.6, and Marital Status=Married then the class of ART Initiation Time of the patient is likely expected to be Delayed for ART (52.0/1.0).

Rule #14: OACD4 = 220.5-252 and Educational Level=tertiary and OA Weight 40.6-57.4 and Marital Status=Separated, then the class of ART Initiation Time of the patient is likely expected to be Delayed for ART (44.52/1.54).

Rule #15: If OACD4= 220.5-252 and OAWHO Stage=I and Age=26-30, then the class of ART Initiation Time of the patient is likely expected to be Delayed for ART (44.52/0.51).

Rule #16: If OACD4= 220.5-252 and Marital Status = Separated and Age=34-38 and OAWHO Stage=I, then the class of ART Initiation Time of the patient is likely expected to be Delayed for ART (22.77/1.02).

4.2.6. Discussion on Major Findings

As the purpose of this research is to identify HIV patient attributes which have strong relationship with ART initiation time and to develop ART initiation time predictive model, the findings are discussed in this section.

The attributes that have strong relation with ART initiation time are determined using Apriori algorithm that identifies association of those attributes with highest probability accuracy of association rule mining. From the eleven HIV patient attributes (Sex, Age, Marital Status, Educational Level, Religion, Family Planning, Occupation, OA Weight, OACD4, OAWHO Stage, and Initiation Time), six attributes are found to have strong association with ART initiation time which are OACD4, WOAHO Stage, Sex, Educational level, OA weight, and Occupation.

The association rule depicts that HIV patients with primary education level who are in Stage IV of WHO Stage has the highest probability of starting ART immediately. Rules generated have indicated that irrespective of any other attributes if OACD4 is below 31.5 or WHO Stage is Stage IV, the patients have to start ART immediately.

Concerning the Sex as a factor to determine ART initiation time the association rules have indicated that Female patients are more likely to start ART immediately than Male patients. Especially Female patients with No Education have to start ART immediately than those with Primary Education, which is an indication that Education level is also another factor in determining ART initiation time.

Accordingly the weight differences between Female HIV patients, those with less weight are expected to take ART Immediately. Female HIV patient with weight in the range of 23.8 to 40.6 are expected to start ART immediately compared to those Female HIV patient with weight in the range of 40.6 to 57.4. Finally concerning Occupation, the association rule mined also has depicted that both Self Employed and Unemployed Female HIV patients are equally expected to start ART immediately.

The researcher has also made efforts to experiment three classification algorithms namely J48, Naïve Bayes and PART. The entire model has generated class predictions for all the predefined classes of the Immediate, Early and Delayed groups with different performance. Out of the selected algorithm, the model selection for this study is done based on the statistical summary obtained from the WEKA machine learning environment and a promising output is also generated using **AdaBoostM1with pruned PART algorithm**. Classification performance has been compared in order to determine optimal statistical algorithms and to extract significant rules for predicting ART Initiation Time of HIV patients.

Since there are many rules, all the obtained rules have not been thoroughly assessed here. Some of the rules that have been discussed hereunder are considered important. However, there may still be other relevant findings in the rules which require time and effort. Therefore these selected rules are again used in prototyping the model to show the applicability of data mining for the identified domain.

4.2.7. 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 enable use of it taking the values of the independent variables to predict ART initiation time 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, OAWHO clinical stage and OACD4 cells count, predict the ART Initiation Time of a patient at high risk i.e. ART Immediate, ART Early and ART Delayed**. For instance, take a patient record with the following details; **Sex = Male, Age = 34-38, Marital Status = Married, Educational Level = No Education, Religion = Orthodox, Family Planning = False, Baseline OAWHO Stage = III, Occupation=Unemployed, OA Weight = 40 and OACD4 =38**. Taking the first rule obtained from the classification algorithm, the patient ART Initiation Time is expected to be below 28 days (ART Immediate). This indicates that a patient beginning with such socio-demographic, clinical and biological parameters needs a due attention to reverse the poor disease prognosis.

The 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 nine predicting variables except one “religion” are selected to enroll in the study. But during knowledge extraction some of the variables predominantly appeared in each of the rules while others occurred less frequently. Therefore, those which occurred frequently are taken as the most predicting variables than the others. Six variables are observed to be much more important in the identified if-then rules. These are; Sex, Age,OACD4 counts, OAWHO Stage, Family Planning, Educational Level, and Occupation, But the rest, Marital Status, religion, and OA Weight are used very less number of times, so that these variables have less predicting capability than the previous seven ones.

1.3. 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 2008. Therefore, only those rules which are suggested to be important by domain experts are placed in to this prototype which means all the rules for predicting ART Initiation Time 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 ART Initiation Time of HIV patients at Optimal Time of ART Immediate, ART Early and ART Delayed of therapy.

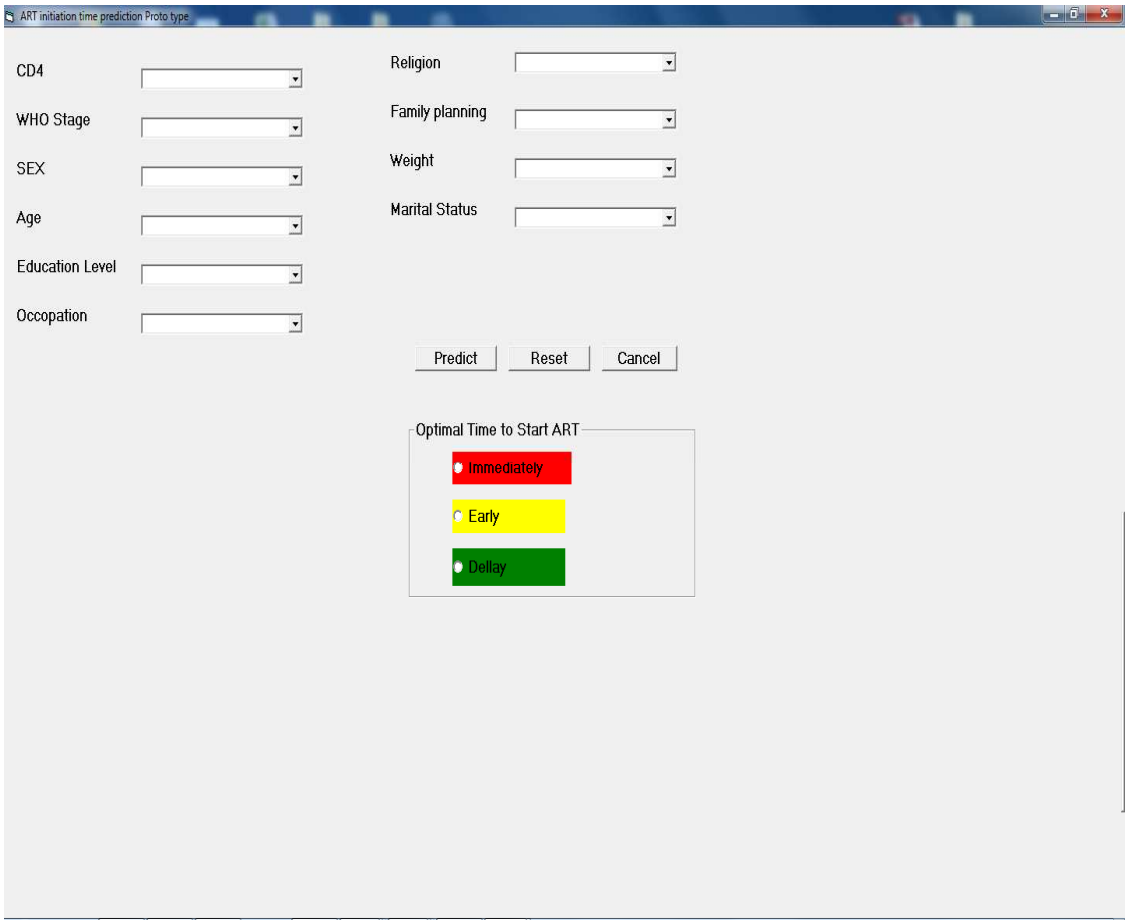


Figure 4.3: Graphical User Interface of the Prototype

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1. Conclusion

Data mining techniques can be applied to ART data and Association Rules help to discover links or patterns in a store of data. It can help to discover interesting associations between data items of patients' records and enable to predict missing or unknown values based on rules mined through the process of association and classification rules mining. The result of ART Initiation Time Prediction in this study is paramount importance for HIV patients, practitioners and policy-makers in determining the most favorable point to begin ART therapy particularly for pre-ART group.

In order to achieve the objective of the study the researcher has been used association rules mining using Apriori algorithm and three commonly used and popular classification algorithms (J48, Naïve Bayes and PART) and long process of data cleansing, data and dimensionality reduction and transformation used it to build the association and prediction models on 11,440 instances and 10 attribute of ART dataset from Addama and Ambo hospital.

The results from association rules using Apriori algorithm showed that attribute values that frequently co-occur together with specific classes of "ART Immediate". Majority of the rule depicted those HIV patients who have no education and those having primary education level, less OACD4 count; OA Weight and Higher OAWHO Stage have the same probability of starting ART immediately. On the other hand HIV patients with higher educational level and in lower WHO stage and higher OACD4 count has the highest probability of starting ART therapy "Early" and "Delayed" in increasing the values of attribute accordingly.

As we have seen the study has shown the necessity to experiment as many classification algorithms as possible before selecting and using a single algorithm for prediction. From classifier models generated for ART Initiation Time, the AdaBoostM1withpruned PART classifier has been selected as a best model based on the accuracy of **95.62%**, mean absolute

error of **0.0301**, FPR and ROC area of **0.027** and **0.994** respectively which is better than the other.

The results of classification rules revealed that OACD4, OAWHO Stage, Sex, age, family planning, are the most determining attributes to predict ART Initiation Time of HIV patients specifically for this study.

5.2. Recommendation

Event thought this research was conducted mainly as an academic Research; its findings have implication far beyond this. It can be used as one component of a decision support system for ART clinic of Ethiopia. The study can contribute a lot for the further studies conducted in the area at Health care where there are huge data amounts.

Particularly on the finding of the research and the cases encountered while conducting the study, the researcher would like to recommend on some point that could benefit the future researcher in the area and the ART clinic that provide therapy.

Trying different algorithms might help to get different results as we tried with Apriori and different classification algorithm. However, the quality of dataset determines the productivity of the data mining process. In the case of Adama and Ambo Hospital ART Clinic, we were able to find some interesting rules between a person's Educational level, OACD4, OA Weight, OAWHO Stage, Occupation and Sex. But we couldn't get as much interesting rules as possible as we have expected which probably emanates from the poor quality of the dataset. The poor quality of the dataset and also inconsistencies observed on the database shows there was poor planning /design for the entry of the data. Therefore, the ART clinic has to evaluate the quality of the ART data, make a thought-through plan to improve quality of data.

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

Further works can be done with the excluded attributes and other techniques Genetic algorithms and Bayesian approach could be encapsulated with Neural Network predictive capability so as to offer more and advantageous results.

Further studies need to be conducted in this field to gather more information on optimal timing of initiation time of HIV patients following TB therapy and factors associated with increased mortality and survival.

REFERENCES

[1]. FHI: HIV voluntary counseling and testing: a reference guide for counselors and trainers. Arlington, USA: Family Health International, Institute for HIV/AIDS.2004.

[2].Federal Democratic of Ethiopia: Country Progress Report on HIV/AIDS Response, April 2012.

[3]. Eyouel T. and Alemayehu W.(2011) Assessment of antiretroviral treatment outcome in public hospitals, South Nations Nationalities and Peoples Region, Ethiopia Ethiop. J. Health Dev. 2011;25 (2).

[4].http://www.aidsinfonet.org.A Project of the New Mexico AIDS Education and Training Center. Partially funded by the National Library of Medicine accessed on May 22, 2012.

[5]. WHO 2006 ART for HIV Infection in Adults and Adolescents: Recommendations for a public Health approach.

[6].The Ethiopian Health And Nutrition Research Institute January 2013 Guidelines for the Implementation of Point-Of-Care Cd4 Testing Technologies in Ethiopia.

[7]. WHO Summary of country profiles for HIV/AIDS treatment scale up 2005.

[8]. Vivek J.andSteven G. Deeks, ; When to Start Antiretroviral Therapy, 2011.

[9].UNAIDS/WHO. AIDS Epidemic Update: UNAIDS/WHO, 2005.

[10].Harvard Medical School report in 2008 on Ethiopia, Kenya and Uganda: Consensus statement on antiretroviral treatment for AIDS in poor countries. Harvard University April 4, 2001.

[11].Matthew S., Stephen B., N. French, Chimota P., Janelisa M. and Eduard E.(2006) Grey nails predict low CD4 cell count among untreated patients with HIV infection in Malawi ISSN 0269-9370.

[12].Degu J. (2007).HIV antiretroviral therapy in Ethiopia Over coming implementation challenges Doctor (PhD) thesis, University of Bergen, Norway.

[13]. Han J., Kamber M. Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann Publishers; 2001.

[14]. George A. Application of Data mining in Medical Applications. Waterloo, Ontario, Canada, 2004.

[15]. Hian Chye K. and Gerald T. Data Mining Application in Healthcare Journal of Healthcare Information Management — Vol. 19, No. 2

[16]. Berry Michael J.A., Linoff Gordon S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Second Edition. Wiley Publishing, Inc.,United States of America. 2004.

[17].Christy, T. Analytical tools help health firms fight fraud. Insurance & Technology, 1997, 22(3), 22-26.

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

[19]. Benko, A. and Wilson, B. Managed Healthcare Executive: Online Decision Support Gives Plans An Edge. , 2003

[20].Simovici, D. Data Mining of Medical Data: Opportunities and Challenges in Mining Association Rules. Retrieved from www.cs.umb.edu/~dsim/papersps/dmmd.pdf, 2012.

[21]. Doddi, S., Marathe, A., Ravi, S. S., & Torney, D. C. Discovery of association rules in medical data. Medical informatics and the Internet in medicine,26 (1), 25-33. Retrieved fromhttp://www.ncbi.nlm.nih.gov/pubmed/11583406,2001.

[22].Fayyad U, Piatetsky-shapiro, G. and Smyth, Padharic. From Data Mining to Knowledge

Discovery in Databases. [database on the Internet];1996.[Access date August 04,2010].

[23].C, K.g;Pedrycz,W.j;Swiniarski,Rw;Kurgan,. Data mining a knowledge discovery approach. Pd; 2007.

[24]. Cios, K. and Kurgan, L. Trends in data mining and knowledge discovery. Springer Verlag, London, UK. , 2005.

[25]. Julio, P. and Adem, K. Data Mining and Knowledge Discovery in Real Life Applications, I-Tech pub. , Vienna, Austria, 2009.

[26]. Factors for Obstetric Fistulae in North-Eastern Nigeria. J Obstet Gynaecol. 2007 Nov; 27(8): 819-23.

[27]. David, H. et. al. Principles of Data Mining. MIT Press, London, UK., 2001.

[28].Adamo, J. M. Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms, Springer-Verlag, New York. 2001.

[29].Shichao Z. and Xindong Wu, Fundamentals of association rules in data mining and knowledge discovery ",John Wiley & Sons, Inc. WIREs Data mining Knowledge Discovery 2011, vol 1 March / April 2011.

[30]. Philippe L., Patrick M., Bonoit V., Stephae L.: "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid ", European Journal of operation research 184 610 – 626, 2008.

[31]. Jiawei H. and Micheline K.: "Data Mining: Concepts and Techniques, 2nd edition". Elsevier publications. 2006.

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

[33]. Bharatheesh T.L.: Predictive data mining for delinquency modeling Bigants Consulting Bangalore, INDIA Iyengar **S.S.** Distinguished Professor Department of Computer Science Louisiana State University Baton Rouge, Louisiana 70803, USA.

[34].Bramer Max. Principles of data Mining. London. Springer -Verlag Limited; 2007.

[35]. Han, J & Kamber: Data mining: concepts and techniques. (2nd ed.). San Francisco: Morgan Kaufmann Publishers, 2006.

[36].Berson, A.; Smith, S.; Thearling, K.: "Building Data Mining Applications for CRM", McGraw-Hill Professional Publishing, New York, USA, (2000).

[37]. Chaudhuri, S.: "Data Mining and Database Systems: Where is the Intersection?", IEEE Bulletin of the Technical Committee on Data Engineering, 21 (1) (1998) 4 - 8.

[38]. Akpınar, H.: "VeriTabanlarında Bilgi KeşfiveVeriMadenciliği", İstanbul Üniv. İşletme FakültesiDergisi, 29, 2000.

[39].K. Gibert et al. / Choosing the Right Data Mining Technique: Classification of Methods and Intelligent…Retrieved from www.iemss.org/iemss2010/index.php?n=Main.Proceedings.

[40]. Joyce J.: Communications of the Association for Information Systems (Volume 8,) 267-296.2002.

[41]. Genadry RR, Creanga AA, Roenneburg ML, Wheeless CR. Complex Obstetric Fistulas. Int J Gynaecol Obstet. 2007 Nov; 99 Suppl 1: S51-6.

[42]. Ng AYJ, M. I. . On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, Neural Information Processing Systems, Ng,A.Y., and Jordan, M. 2002.

[43]. Ian Witten H & Eibe F. Data mining: practical machine learning tools and techniques. Second edition.San Francisco: Morgan Kaufmann Publishers, 2005.

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

[45]. Vinterbo, S. A. (1999). Predictive Models in Medicine: Some Methods for Construction and Adaptation. Norwegian University of Science and Technology, Oslo, Norway.

[46] Selam, A. (2011).Predicting the Occurrence of Measles Outbreak in Ethiopia Using DM Technology.MSc. Thesis, Addis Ababa University, Ethiopia.

[47]. Amir, F. and Shahram, J. (2011).An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network. International Journal of Advanced Science and Technology. Vol. 34, Shiraz University, Shiraz, Iran.

[48]. S. P. Deshpande and V. M. Thakare). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS) Volume 1, Number 1: 445-463, 2010.

[49]. Two Crows Corporation: Introduction to Data Mining and Knowledge Discovery. 3rd Ed .Two Crows Corporation. 500 Falls Road, Potomac, USA, 2005.

[50].Wipada Chanthaweethip, SumantaGuha, Temporal Data Mining and Visualization for Treatment Outcome Prediction in HIV Patients, Asian Institute of Technology, Pathumthani 12120, Thailand, 2012.

[51]. Madigan, E.A., Curet, O.L., & Zrinyi, M.(2008) .Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns. Human Resources for Health, 6(2):1-6.

[52].Abraham, T.(2005). Application of data mining technology to identify determinant risk factors of HIV infection and to find their association rules: the case of center for disease controls and prevention (CDC).M.Sc. thesis, Addis Ababa University, Addis Ababa ,Ethiopia.

[53].Vararuk, A., Petrounias, I. & Kodogiannis, V.(2008).Data mining techniques for HIV/AIDS data management in Thailand. Journal of Enterprise Information Management, 21 (1):52-70.
[54]. 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.

[55].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.

[56]. Rosma M, Sameem A, Basir A, Adeeba K, and Annapurni K. The Prediction of AIDS Survival: A Data Mining Approach. Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering.
[57]. Larose Daniel T. Discovering Knowledge in Data - An Introduction to Data Mining. New Jersey, USA: John Wiley & Sons Inc; 2005.
[58]. Behailu G. Constructing a Predictive Model for Determining CD4 Status of Patients Following Art: The Case of Jimma and Bonga Hospitals [Unpublished MSc Thesis]. Addis Ababa University: School of Information Science and School of Public Health; 2012.
[59]. Minale T. Application of Data Mining Techniques to Predict Urinary Fistula Surgical Repair Outcome [Unpublished MSc Thesis]. Addis Ababa University: School of Information Science and School of Public Health; 2012.

Appendix A: Attribute Ranking for the ART initiation Time Prediction

=== Run information ===

Evaluator: weka.attributeSelection.GainRatioAttributeEval

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

Relation:000_afterprocessed-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-B10-M-1.

Instances: 11440

Attributes: 11

Sex, Age, MaritalStatus, EducationalLevel, religion, FamilyPlanningYN

Occupation, OAWeight, OAWHOStage, OACD4, initiationtime

Evaluation mode: evaluate on all training data

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

Search Method:

Attribute ranking.

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

Gain Ratio feature evaluator

Ranked attributes:

0.358991 10 OACD4

0.115544 1 Sex

0.065414 9 OAWHOStage

0.030865 4 EducationalLevel

0.021747 8 OAWeight

0.015954 7 Occupation

0.007501 6 FamilyPlanningYN

0.001146 3 MaritalStatus

0.000888 5 relegion

0.000489 2 Age

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

Appendix B: Sample Output of pruned J48 Selected Scheme

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.5 -M 2

Relation:     000_after processed-weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-

Instances:    11440

Attributes:   11

=== Summary ===

Correctly Classified Instances       10929               95.5332 %

Incorrectly Classified Instances      511                4.4668 %

Kappa statistic                          0.8824

Mean absolute error                      0.0348

Root mean squared error                  0.1629

Relative absolute error                 13.7851 %

Root relative squared error             45.8844 %

Total Number of Instances            11440

=== Detailed Accuracy By Class ===

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

                 0.893    0.026     0.832    0.893    0.861     0.95     Delayed

                 0.793    0.017     0.846    0.793    0.818     0.931    Early

                 0.987    0.031     0.991    0.987    0.989     0.985    Immediate

Weighted Avg.    0.955    0.029     0.956    0.955    0.955     0.975

=== Confusion Matrix ===

   a    b    c   <-- classified as

 1274  117   35 |    a = Delayed

  199  940   47 |    b = Early

   59   54 8715 |    c = Immediate
```

Appendix C: Sample Output of Unpruned J48 Selected Scheme

```
=== Run information ===

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

Relation:     000_after processed-weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last- Instances:    11440

Attributes:   11

=== Summary ===

Correctly Classified Instances       10929               95.5332 %

Incorrectly Classified Instances      511                4.4668 %

Kappa statistic                       0.8824

Mean absolute error                   0.0348

Root mean squared error               0.1629

Relative absolute error              13.7851 %

Root relative squared error          45.8844 %

Total Number of Instances            11440

=== Detailed Accuracy By Class ===

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

                 0.893    0.026     0.832     0.893     0.861      0.95     Delayed

                 0.793    0.017     0.846     0.793     0.818      0.931    Early

                 0.987    0.031     0.991     0.987     0.989      0.985    Immediate

Weighted Avg.    0.955    0.029     0.956     0.955     0.955      0.975

=== Confusion Matrix ===

   a    b    c   <-- classified as

 1274  117   35 |    a = Delayed

  199  940   47 |    b = Early

   59   54 8715 |    c = Immediate
```

Appendix D: Sample Output of AdaBoostM1with + pruned J48 AdaBoostM1with + pruned J48

```
=== Run information ===

Scheme:    weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 -- -C 0.5 -M 2

Relation:     000_after processed-weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last

Instances:    11440

Attributes:   11

=== Summary ===

Correctly Classified Instances       10925               95.4983 %

Incorrectly Classified Instances      515                4.5017 %

Kappa statistic                          0.8812

Mean absolute error                      0.0306

Root mean squared error                  0.1655

Relative absolute error                 12.1335 %

Root relative squared error             46.6066 %

Total Number of Instances            11440

=== Detailed Accuracy By Class ===

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

                 0.874     0.023     0.846     0.874     0.86      0.984    Delayed

                 0.81      0.019     0.834     0.81      0.822     0.977    Early

                 0.988     0.037     0.989     0.988     0.988     0.996    Immediate

Weighted Avg.    0.955     0.033     0.955     0.955     0.955     0.993

=== Confusion Matrix ===

   a    b    c   <-- classified as

 1246  135   45 |    a = Delayed

  173  961   52 |    b = Early

   54   56 8718 |    c = Immediate
```

Appendix E: Sample Weka Output of the selected Model

```
=== Run information ===

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

Relation:     000_after processed-weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-

Instances:    11440

Attributes:   11

              Sex                     Age                     Marital Stattus

              EducationalLevel        Religion                FamilyPlanningYN

              Occupation              OAWeight                OAWHOStage

              OACD4                   initiationtime

Test mode:    10-fold cross-validation

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

AdaBoostM1: Base classifiers and their weights:

=== Summary ===

Correctly Classified Instances       10939               95.6206 %

Incorrectly Classified Instances       501               4.3794 %

Kappa statistic                          0.8847

Mean absolute error                      0.0301

Root mean squared error                  0.1645

Relative absolute error                 11.9231 %

Root relative squared error             46.33   %

Total Number of Instances            11440

=== Detailed Accuracy By Class ===

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

                 0.877     0.024     0.841     0.877     0.859      0.987    Delayed

                 0.814     0.018     0.836     0.814     0.825      0.979    Early

                 0.988     0.029     0.991     0.988     0.99       0.997    Immediate

Weighted Avg.    0.956     0.027     0.957     0.956     0.956      0.994

=== Confusion Matrix ===

   a    b    c   <-- classified as

 1250  140   36 |    a = Delayed

  181  965   40 |    b = Early

   55   49 8724 |    c = Immediate
```