ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

JOINT PROGRAM BETWEEN FACULTY OF INFORMATICS AND MEDICAL FACULITY

DEPARTEMNT OF HEALTH INFORMATICS

APPLICATION OF DATA MINING TECHNIQUES ON ANTIRETROVIRAL THERAPY (ART) DATA: THE CASE OF ADAMA AND ASELLA HOSPITALS

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABAB UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN HEALTH INFORMATICS

BY

TEKLU URGESSA ABEBE

JUNE, 2010

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

JOINT PROGRAM BETWEEN FACULTY OF INFORMATICS AND MEDICAL FACULITY

DEPARTEMNT OF HEALTH INFORMATICS

APPLICATION OF DATA MINING TECHNIQUES ON ANTIRETROVIRAL THERAPY (ART) DATA: THE CASE OF ADAMA AND ASELLA HOSPITALS

BY

TEKLU URGESSA

June, 2010

NAME AND SIGNATURE OF MEMBERS OF THE EXAMINATION BOARD

Chairman, Department Examination Board: Dr. Million Meshesha _____

Examiner: Dr. Dereje Teferi _____

Advisor: Dr. Sebsibe Haile/Mariam _____

DEDICATION

This Paper is dedicated to My Mother, W/ro Sinke Mullata.

DECLARATION

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

_____

Teklu Urgessa

June 2010

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

_____

Dr. Sebsibe Hailemarriam

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. Sebsibe Hailemariam for his constructive and uninterrupted comments and guidance through out the research work. Had it not been for his strict follow up and continued comments, this research would have not been realized.

I would like to acknowledge the management and database clerks of Adama and Asella hospitals for their cooperation in providing me with data and some explanations regarding the data. Again commenting on the knowledge obtained from the research finding.

I would like to thank my Love Hawi Abdurkadir for the mental peace and love she gives me in doing this research.

I am also thankful to my babies Bontu Teklu and Fraol Teklu, for whom, I avoided sense of carelessness in my life and being source of my happiness.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **III**
TABLE OF CONTENTS ..... **IV**
LIST OF ABBREVIATIONS ..... **VII**
LIST OF TABLES ..... Error! Bookmark not defined.
LIST OF FIGURES ..... **X**
ABSTRACT ..... **X**
CHAPTER ONE **1**
INTRODUCTION **1**
1.1 Background 1
1.1.1. HIV/AIDS Burden and Actions taken in Ethiopia 1
1.1.2. ART and Guidelines for its Treatment 2
1.1.3 Relevance of Data mining to ART 5
1.1.4 ART Care in Ethiopia 6
1.2 Statement of the Problem 10
1.3 Objectives the study 12
1.3.1 General Objective 12
1.3.2 Specific Objectives 12
1.4. Research Methodology **13**
1. 4.1 Data Mining Modeling 13
1. 4.1.1. Business/Problem Understanding 13
1. 4.1.2. Data collection and Understanding 13
1.4.1.3 Data Preparation and Preprocessing 14
1.4.1.4 Model Building 14
1. 4.1.5 Analysis and Evaluation of the models 14
1.4.2. Tools 16
1.4.3 Literature Review 16
1.4.4. Ethical Considerations 16
1.4.5. Dissemination of Results 17

1.5 Scope and Limitation of the study 17
1.6 Organization of the Thesis 18
CHAPTER TWO 19
DATA MINING TECHNOLOGY AND ITS APPLICATION IN HEALTH CARE 19
2.1 What is Data Mining? 19
2.1.1 Data Mining Classification 21
2.1.2 Association Rules 24
2.1.3 Clustering 25
2.1.5 Data Mining Processes 27
2.2 Application of Data Mining in Health Care 31
2.2.1 Related Works 33
CHAPTER THREE 37
DATA PREPROCESSING AND MODEL SELECTION 37
3.1 Data Source Description 37
3.2 Descriptive Statistical Summary of Attributes 42
3.2.1 ARTStage 42
3.2.2 Marital Status 42
3.2.3 Religion Attribute 43
3.2.4 Sex Attribute 43
3.2.5. Educational Level 44
3.2.6 OAWHO Attribute 44
3.2.7 Occupation Attribute 45
3.2 .9 Functional Status Attribute 46
3.2.10 Year Attribute 46
3. 2.11 Termination Attribute 47
3. 2.12 Month-Year 48
3.2.13 OACD4 Attribute 48
3.2.14 Age Attribute 49
3.3 Data Cleaning 50
3.3.1 Handling Missing Values 50
3.3.2 Handling Outliers Values 51

3.3.3 Handling Noisy Values 52
3.3.4 Data Transformation and Reduction 52
3.4 Model Implementation and Experimentation 56
3.4.1 Decision Tree 57
3.4.2 Implementation of Association Rule 63
CHAPTER FOUR 66
MODEL BUILDING 66
4.1 Attribute Ordering 66
4.2 Building Classification Models 67
4.2.1 Selection of Validation Method for Decision Tree Models 69
4.2.1 Building Binary Decision Tree 70
4.2.2 Modeling Generalized Decision Tree 71
4.3: Building Association Rule Model 73
CHAPTER FIVE 76
EXPERIMENT AND ANALYSIS OF CLASSIFICATION MODEL 76
CHAPTER SIX 96
EXPERIMENTS AND ANALYSIS OF ASSOCIATION 96
CHAPTER SEVEN 113
CONCLUSIONS AND RECOMMENDATIONS 113
6.1 Conclusions 113
6.2 Recommendations 117
References 118

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ARV	Antiretroviral
CD4	Clustered Differentiation 4
CRISP-DM	Cross Industry Standard for Data Mining
EIS	Executive Information System
FPR	False Positive Rate
HAART	Highly Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
IDU	Injection Drug Use
IRM	International Resource Management
KDD	Knowledge Discovery in Database
MAC	Millennium AIDS Campaign
MOH	Ministry of Health
RDBMS	Relational Database Management Systems
ROC	Receiver Operating Characteristics
TDIDT	Top Down Induction of Decision Tree
TPR	True Positive Rate
VCT	Voluntary Counseling and Testing
WHO	World Health Organization
UNAIDS	Joint United Program Nations Program
NAC	National AIDS Counsel
HAPCO	HIV/AIDS Prevention and Control Office

LIST OF TABLES

Table 1.1: Number of patients on ART in Ethiopia by Region 7
Table 1.2: Variables Representing Confusion Matrix of the Models 14
Table 3.1: Source and Number of Records 38
Table3.2: Description of the Attributes 39
Table 3.3: Statistical Summary of ART Stage Attribute 42
Table 3.4: Statistical summary of Marital Status Attribute 42
Table 3.5: Statistical Summary of Religion Attribute 43
Table 3.6: Statistical Summary of Sex Attribute 44
Table 3.7: Statistical summary of Educational level attribute of the patients 44
Table 3.8: Statistical summary of OAWHO attribute 45
Table 3. 9: Statistical Summary Measure for the Occupation Attribute 45
Table 3.10: Statistical Summary of Functional Status Attribute 46
Table 3.11: Statistical Summary of Year Attribute 47
Table 3.12. Statistical summary of the termination status 47
Table 3. 13: Statistical Summary of OACD4 Attribute 48
Table 3.14: Statistical Summary of Age Attribute 49
Table 3.15: Summary of Modal values for Nominal Attributes **with missing Values 51**
Table 3.16: Discretized result of Age attributes 54
Table 3.17: Result of Discretization for OACD4 55
Table 3.18: Description of J48 classifier Parameter Options in Weka 62
Table 4.1: Samples of Training Dataset and Corresponding Performance of Classifier 70
Table 4.2: Meanings of the Parameters for Association Rule 75
Table 5.1: Summary of Scenario #1 78
Table 5.2: Performance Summary of Scenario #2 80
Table 5.4: Summary of performance scenario #3 82
Table 5.5: Summary of performance of scenario #4 83

Table 5.6: Summary of performance of scenario # 5 84
Table 5.7: Summary of performance of scenario #6 87
Table 5.8: Summary of performance of scenario #7 87
Table 5.9: Summary all measures of performance for all models 90
Table 6.1: Experiments made for association rule mining 96
Table 6.2: Number of Rules for Each Minimum Thresholds 110

LIST OF FIGURES

Fig. 2.1 Data Mining Processes.....29

Fig. 2.2: Phases of CRISP-DM 30

Fig. 3.2 Instances Matching Antecedent, Consequent and both 64

Fig. 4.1: Learning Curve for Training dataset.....70

Fig. 4.2: Weka Explorer Window.....72

Fig. 4.3: Weka Parameters Window for J48 Classifier.....72

Fig. 4.4: Weka Apriori Window for parameter setting.....74

Fig. 5.1: ROC Area curve for Scenario#1.....79

Fig. 5.2: Model structure of General Decision Tree with Pruning and with all attributes.....86

Fig. 5.2: Graph for Tree Complexity.....91

Fig. 5.3: Time Trends for all models.....91

Fig. 5.4: Comparison by performance for all models.....92

Fig. 5.5 comparison of the models by accuracy measures 92

ABSTRACT

Human Immunodeficiency Virus/ Acquired Immunodeficiency Syndrome (HIV/AIDS) is of global as well as national concern today as it affects all people of the world regardless of sex, age, educational status, race and color. When we come to Sub-Saharan African region in general and Ethiopia in particular, the situation is even more worsening and needs special attention. Today more than 1 million people are living with HIV/AIDS in Ethiopia. The country has made a lot of efforts towards preventing and controlling of the disease. As a result, hundreds of thousands of people come to health facilities to get Counseling and testing services through Voluntary Counseling and Testing (VCT) and Antiretroviral Therapy (ART) programs.

A lot of demographic and Clinical data is recorded about individuals taking the services. As these data is getting larger and larger, it is highly likely that there will be hidden, implicit and non trivial knowledge within the data, which might not be obtained by the traditional statistical analysis as well as report and query based database functionalities. There are various evidences that Data Mining (DM) helps the health care system to extract non-trivial and hidden knowledge which exists within the large volume of demographic and clinical data captured during the provision of services and that this knowledge is helpful for health administrators to target resources in the right directions for preventive and controlling activities, and clinicians to give safe and right treatment and saves humans' lives.

Therefore; the main objective of this research was to see the applicability of data mining techniques on ART data collected at facility level by taking the case of Adama and Asella Hospitals ART databases to identify important patterns related to determinant attributes and their values for Termination/ Continuity behavior of patient on ART care service. Various data preprocessing activities were made to come up with the dataset ready for model building. The researcher selected two DM functionalities (Classification and Association rules mining). Decision tree classification with J48 implementation with eight scenarios was experimented. Thirteen experiments with different parameters were made for association rule mining. Evalution of the models was performed by using for each DM functionality and scenarios used to model the dataset. Analysis of the model was made based on different criteria mainly using confusion matrix, accuracy measures,

time of execution and tree complexity for decision tree classification models and number of rules generated, support and confidence for each scenario of the association rule

The research showed encouraging results; that data mining techniques are of high potential in predicting determinant factors/attributes for termination/continuity behavior of ART care by the patients.

Finally hidden patterns (knowledge) were extracted that will provide certain decision support information for concerned bodies, for ART programs intervention. To mention few, the result showed for example that those patients who were on ART stage and whose Functional status is bedridden and the year in which they began the service is before 1999 E.C are at high risk of terminating the ART care. Those patients whose ART stage is on ART, and whose functional Status is Ambulatory, and if they started the service before 1999 and their age is above 18 years then they have high chance to terminate the ART care. The study also showed certain hidden information that young people whose age is less than 18 years; have high chance of staying longer in ART care service. Patients terminate the service in shorter time at Asella hospital than at Adama hospital. Those who are jobless have high chance to stay in the care. The reason (s) for these hidden patterns is left open for future researches works. From comparisons done among the experimentations made, it was learned that those data mining techniques, which were experimented for this research are applicable on the ART dataset of the cases under investigation in general but generalized decision tree with pruning outperformed for classification purpose on the dataset in terms preciseness, providing general insight, Performances and accuracy measures with fair execution time and providing best interpretable patterns. Many association rules were obtained with minimum support of 30% and confidence 50% had provided optimum rules with acceptable patterns.

CHAPTER ONE

INTRODUCTION

1.1 Background

1.1.1. HIV/AIDS Burden and Actions taken in Ethiopia

Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) pandemic continues to spread worldwide. Very large number of people (range 34.6-42.3 million) were living with the virus, which killed about 3 million in 2003, and over 20 million since the first case of AIDS were identified in 1981. Sub Saharan Africa, with only 10% of the total world population, is carrying the burden of 80% of the world HIV infection and AIDS cases (UNAIDS, 2004).

Ethiopia is one of the countries in Sub-Saharan -African region which is hard hit by HIV/AIDS. In Ethiopia, the first HIV/AIDS case was reported in 1986. The disease spread was so fast and in 2003, there were more than 1.5 million people living with HIV/AIDS in Ethiopia of which 817,000 were women and 96,000 were children under the age of 14 (Ethiopian Ministry of Health Report, 2004). This data suggested that in urban Ethiopia, the HIV epidemic spread rapidly and reached a plateau around the mid-1990s and stabilized afterwards (UNAIDS, 2006).

In response to the HIV/AIDS epidemic, the Ethiopian government issued HIV/AIDS policy in 1998; established in 2000 the National AIDS Council (NAC), under the office of the Prime Minister; The National HIV/AIDS Prevention and Control Office (HAPCO) was established by Proclamation in July 2002 as the executive arm of the NAC; and Antiretroviral Therapy (ART) was first publicly available in Ethiopia in 2003, and the government launched free access to ART nationwide in 2005 (MOH, 2006)

Over the past decade, access to ART has grown at an unprecedented rate in sub-Saharan Africa; with tremendous health gains observed among those utilizing these services (Kruse R. et al, 2009). HIV/AIDS Prevention and Control Office (HAPCO) 2007 reports indicated that an ART service was limited to hospitals, but since June 2006 the service has expanded to include health centers. As a result, hundreds of thousands of people come to health facilities to get Counseling and Testing through Voluntary Counseling and Testing (VCT) and to get treatment through Antiretroviral Therapy (ART) programs.

1.1.2. ART and Guidelines for its Treatment

Antiretroviral drugs are medications for the treatment of infection by retroviruses, primarily HIV. When several such drugs, typically three or four, are taken in combination, the approach is known as Highly Active Antiretroviral Therapy (HAART). The American National Institutes of Health and other organizations recommend offering antiretroviral treatment to all patients with AIDS. Because of the complexity of selecting and following a regimen, the severity of the side-effects and the importance of compliance to prevent viral resistance, however, such organizations emphasize the importance of involving patients in therapy choices, and recommend analyzing the risks and the potential benefits to patients without symptoms (WHO, 2003)

There are some guidelines to give ART treatment services. 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 (that is, developing nations), HIV-infected adults and adolescents should start ART when HIV infection has been confirmed and one of the following conditions is present:

1. Clinically advanced HIV disease;
2. WHO Stage IV HIV disease, irrespective of the CD4 cell count;
3. WHO Stage III disease with consideration of using CD4 cell counts less than 350/µl to assist decision making;
4. WHO Stage I or II HIV disease with CD4 cell counts less than 200/µl.

It is found important to the researcher to define the jargons CD4 Cells, CD4 counts and WHO clinical stages for ART as background since understanding the criteria is difficult without it.

CD4 cells are those cells, which have molecules called CD4 on its surface. These "helper" cells initiate the body's response to invading micro-organisms such as viruses. HIV is a retrovirus, meaning it needs cells from a "host" in order to make more copies of itself (replication). In the case of HIV, CD4 cells are the host cells that aid HIV in replication. HIV attaches to the CD4 cells, allowing the virus to enter and infect the CD4 cells, damaging them in the process. The fewer functioning CD4 cells, the weaker the immune system and therefore the more vulnerable a person is to infections and illnesses (WHO, 2006).

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 resource poor countries 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. The following are the WHO clinical stages of HIV progression (WHO, 2006):

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:-

- 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 lasting for longer than one month
- Unexplained persistent fever, either intermittent or constant
- Persistant 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 infections (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.

There are also several concerns about antiretroviral treatments even though the above criteria are considered. The drugs can have serious side-effects. Treatment can be complicated, requiring patients to take several pills at various times during the day, although treatment regimens have been greatly simplified in recent years. If patients miss doses, drug resistance can develop. Also, providing anti-retroviral treatment is costly and resource-intensive, and the majority of the world's infected individuals cannot access treatment services. Research to improve current treatments includes decreasing side effects of current drugs, further simplifying drug regimens to improve adherence, and determining the best sequence of regimens to manage drug resistance is going on. However; discontinuation/ termination of ART care follow up by the patient is an other issue due to various reasons which needs attention and research on factors affecting the continuity (Morris, J. et. al. 2007). This is exactly what this research is intended to investigate using data mining techniques.

1.1.3 Relevance of Data mining to ART

During ART services delivery to HIV patients, large volume of data is collected. The data collected over time from the patients, can tell so much if appropriately analyzed as to what factors affect the continuity of ART care use. However as data is growing larger and larger, the data carry a lot of implicit knowledge that can not be accessed with simple statistical analysis methods. Moreover, the larger the databases size, the more difficult to utilize the data for use in intelligence based decision support system. Hence it becomes apparent that traditional query and report based system as well as statistical analysis techniques cannot help for the purpose. The most appropriate technique to discover and manage knowledge implicitly hidden inside the collected data is data mining (Abdul-Kareem, S. et. al. (2000). Abdul-Kareem stated importance of data mining as:

> *Data mining combines techniques from database research, artificial intelligence and statistics and used in a variety of domains. The concept of data mining is becoming increasingly popular as an information management tool as it reveals the knowledge implicit in the data that has been collected by various application areas over period of times*

The health care system issues of data overload and lack of appropriate analysis tools to come up with decision support information has also been raised in different literatures; among which, Kaur H. and Wasan S, (2006) stated it as "The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor". They extended their explanation as "There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Valuable knowledge can be discovered from application of data mining techniques in healthcare system"

Data mining approaches can be applied to clinical databases to assist with decision support. Clinical health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care (Piatetsky-Shapiro et. al. 1991).

1.1.4 ART Care in Ethiopia

According to the most recent estimates, about 1 million people (2.2% of the adult population) were living with HIV in Ethiopia in 2008. In the same year, approximately 290,000 people needed ART (Ethiopian Ministry of Health 2007). To respond to the treatment needs of people living with HIV/AIDS, the National Antiretroviral Drugs Policy was developed in 2002, and the first treatment guideline for adults and adolescents was issued in 2003 and revised in 2007 (Ethiopian Ministry of Health 2007). A fee-based ART program was officially started in 2003. Moreover, a number of initiatives have been

undertaken to expand the availability of ART in Ethiopia, including those by the Global Fund, PEPFAR, the Ethiopian North American Health Professionals Association, the Clinton Foundation, and the Ethiopian Red Cross Society. As a result, a free ART program was launched in early 2005. Under the guidance of the strategic plan for the multi-sectoral response, 2004–2008 PEPFAR (The United States President's Emergency Plan for AIDS Relief 2007) and the road map for accelerated access to ART, 2004–2006 and 2007–2008/10 (Ethiopian Ministry of Health 2004), the ART roll-out plan has been implemented. Consequently, ART services have been decentralized and have been available in both health centers and hospitals since August 2006.

Despite the many challenges, ART scale up has recorded the greatest achievement over the last few years. The service has been expanded from only three health facilities in 2005 to 835 governmental 65 private in 2008. The number of people ever started on ART has also shown an unprecedented increase during the same period from 900 in 2005 to 180,447 by end of December 2008. The total number of people on ART and Pre ART by regions in Ethiopia in 2008 as reported in Millennium AIDS Campaign (MAC) document 2007 is as in Table 1.1.

Table 1.1: Number of People on ART in Ethiopia by Region

Region	Pre ART	On ART	Gov. Facilities	Private Facilities
Tigray	29,395	14436	65	5
Afar	5,279	361	11	1
Amhara	98,697	51317	231	18
Oromiya	76,120	37163	167	13
Somoli	5,798	3163	14	1
Benishangul	4,666	1574	7	1
SNNPR	44,144	19483	88	7
Gambel	1,582	910	4	1
Harari	2,793	1249	6	1
Addis Ababa	81,351	43907	197	16
Dire Dawa	4,980	2981	13	1
Uniformed	9,919	7,039	32	0
Total	364,724	183583	835	65

Adama and Asella Hospitals are zonal referral hospitals in Ethiopia which deal with large volume of clinical data in general and of HIV/AIDS in particular. Both hospitals have ART database but the data in the database is simply used for ordinary reporting purpose by using database query and report functionalities. The number of records in both databases is around 18740 today. Currently the facilities register the patient who know their HIV status and need to begin the service. The patient is appointed for follow up. The patient stays in care as pre ART until they become eligible to take the antiretroviral (ARV) drug. The stage is registered as "IN" to mean 'in care' if the patient is on Pre ART. Treatment for opportunistic infections is given until they become eligible for ARV. For eligibility, the baselines are WHO stage of AIDS and CD4 count of individual patients. The ART stage changes to EL which stands for Eligible when the patient is eligible to take the ARV drug i.e when they met the criteria put by WHO as mentioned above in section 1.1.2. The patient is given Adherence counseling to make the patient ready for taking ARV. When the patient is ready to take the ARV drug as per the adherence counseling, the patient ART stage is updated into ER (Eligible and Ready). The ART Stage is set to OA (On ART). There termination status on different appointments is also registered. Other clinical and demographic data of the patient is also recorded. The database clerks can generate reports, formulate simple query and update the patient data as required. The researcher's interest in this research is there might be hidden patterns or relationships within the data in the ART database that might not be obtained by using ordinary database queries and reports regarding what factors affect the continuity/termination of ART use by the patients even though it is possible to get total number of patients who terminated the ART care use.

This study is designed to see the applicability of data mining techniques to identify important patterns/relationships in ART database focusing on what factors might affect the continuity/termination of ART care use by the patients; once they started it in Ethiopian context by taking Adama and Asella Hospitals as the cases.

Chausa, P. et al. (2009) conducted a research through data mining techniques to the analysis of the data collected since 1981 by the Infectious Diseases Unit of a Hospital Clinic in Barcelona, Spain. They stated that although the Highly Active Antiretroviral

Therapy (HAART) reduced the number of AIDS cases since 1996 by significantly increasing the disease-free survival time, the therapy failure rate is still high due to HIV treatment complexity. Their study was aimed to understand changes in the outcomes of HIV infected patients, who use the HAART. In their study, they stated that their interest was looking for two types of treatment failures: viral failure and toxic failure, corresponding to events of clinical interest to assess the treatment outcomes from the ART data. They suggested that the analysis allowed them to extract different typical patterns related to each period and to meaningfully interpret the previous and current behavior of HIV Therapy. This research is designed to see different demographic and clinical characteristics in relation with the termination status of the ART care use by the patient using data mining techniques.

1.2 Statement of the Problem

The underlining problem which necessitated this research is the ever growing pandemic of HIV/AIDS in the world in general and in Ethiopia in particular and complexity of the problems in stopping and controlling it. The problem needs due attention and become area of research for many scholars. Tremendous efforts are being made in preventing and controlling it. One of the efforts made in minimizing death and promotes disease free survival of people infected by HIV is the free availability of ART care. But there are so many factors complicating the successful delivery of ART care services at the facility levels. It is reported by facilities that patients terminate the care but it believed that if patients follow it though out their life, it will bring better health to them and enable them live longer. To know these exacerbating factors, it is important to see the patterns within the data collected during the service delivery. On the other hand, as the data collected at the facility level is increasing to a larger volume, it needs modern (state-of –the-art) analysis technology to extract hidden patterns, this is possible using data mining techniques (Chausa, P. et al. 2009).

According to Vararuk et. al. (2008), data collection for HIV/AIDS related intervention typically involves thousands of data items. From all this data, the health care provider must combine data elements for decision making related to HIV/AIDS and plan for intervention. ART care service is one of the services and most critical. It is important to assess factors that affect continuity of the patient on the service once he/she has started taking it at facility level. ART care services are expanding in every nation of the world of which Ethiopia is in mention nowadays. Not only expanding services but also studying what factors might affect the success of the program from the data collected during service delivery through powerful analysis tools has paramount importance to maintain continuity and promote good health of HIV infected people by preventing opportunistic infections and hence delaying death. The data may be too much to be easily analyzed by the classical statistical methods as well as simple queries of database management system. Data mining is a solution for discovering hidden but important patterns from large volume of data collected over time.

When we come to our country; data mining research on health care system in general and HIV/AIDS in particular is almost non existent but it is believed that if data mining applied in this area, it might be critically important in revealing non trivial decision support information for both health care administration and treatment. Therefore; the researcher is motivated to see the potential applicability of data mining techniques on ART data in the Ethiopian context by taking Adama and Asella hospitals as cases. Adama and Asella Hospitals, which are taken as cases of the research provide VCT and ART services and deals with large volume of HIV/AIDS data, which should be manipulated further to derive new knowledge. They have ART databases on which they handle data about the patients. The database is currently used for ART care services (follow up) and generation of reports up on need. But it is the researcher's believe that if data mining techniques are applied on this data, there might be new pattern or relationship among attributes of database and the knowledge obtained is used to improve the success and efficiency of the ART care in these facilities or it may pave the way for conducting further investigation to find out what factors might affect the continuity of ART care use by the patients. The study is intended to apply data mining techniques on ART database of the selected Hospitals to identify patterns associated to attribute that stands for termination status of the patient. The dataset consists of the attribute 'Item No', 'Card No', 'Registration Date', Eligible Date', Eligible and Ready Date', 'On ART Date' "Termination", "Age", "Sex", "Educational Level", "Religion", "ART Stage", 'Functional Status', "WHO Stage of ART", "Marital Status", "CD4 count" and "Occupation". The dependent variable is "Termination Status" and the rest of the attributes are independent variables.

The outcome of the research has contribution in the identification of important patterns within ART data of the taken cases and helps the decision makers to base their intervention for ART Care. It can also be used as the corner stone for further studies in the area. It enables us to evaluate the efficiency of data mining techniques in the area of investigating clinical data particularly in the area of HIV/AIDS.

1.3 Objectives the study

1.3.1 General Objective

The main objective of the study is to see the applicability of data mining techniques on ART data by taking Adama and Asella Hospitals as cases to come up patterns that affect continuity/termination behavior of patients once registered for the service.

1.3.2 Specific Objectives

The specific objectives are to:

- Review different literatures that can support the study in the area of applying data mining techniques on health care in general and ART in particular.
- Selecting the ART database for data collection
- Collect and analyze working data which is relevant to the data mining problems
- Pre-process and preparing the raw data into a suitable dataset for experiment for the data mining functionalities
- Selecting the models to be implemented
- Building data mining models on the pre-processed dataset
- Analyze the results of the models in terms of various performance matrices and interestingness measures of the association rules
- Concluding and recommending on the basis of possible knowledge obtained from the research regarding the applicability of DM techniques on ART dataset in discovering factors affecting continuity or termination behavior of ART care use by patients.

1.4. Research Methodology

1. 4.1 Data Mining Modeling

The Data Mining Modeling methodology adopted in this research is CRISP_DM (Cross-Industry Standard Process for Data Mining), which involves Business understanding, data understanding, data preparation and pre-processing, model building, evaluation and deployment steps. In this section, the researcher tries to briefly discuss the steps in this methodology.

1. 4.1.1. Business/Problem Understanding

Business/problem understanding in the CRISP_DM is critical DM process research activity as understanding the research area/Problem domain. It focuses on understanding the research objectives and requirements and goals as stated by Chapman, P. et. al., (2000). Hence, in this research, the effort is made to formulate the research problem in the way it is addressed by data mining techniques. Objective of the research was also set as well as the requirements to attain the main objective specified. Data mining goal is also speculated.

The purpose of the research is to see the applicability of data mining techniques in discovering important patterns in ART data focusing on factors affecting continuity/termination behavior of the ART care services once they are registered for it on the taken cases.

1. 4.1.2. Data collection and Understanding

In this phase, data is collection of original data is made. Activities in order to get familiar with the data were took place. Efforts to identify data quality problems were made, which helped the researcher to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. Data is collected for the ART service continuity/Termination study from Adama and Asella Hospitals ART databases. A total of 18,740 records of the ART databases of both hospitals are used for the research. The dataset consists of a total of 17 attributes. These attributes are: ‘Item Number’, ‘Card Number’, ‘Registration date’, Eligible date’, Eligible and Ready date’, ‘On ART Date’,

‘Sex’, ‘Age’, Functional Status ,‘Termination’, ‘Religion’, ‘Educational Level’, ‘ART Stage’, ‘Occupation’, ‘Marital Status’, ‘OAWHO’ (WHO Stage) and ‘OACD4’ (CD4 count). The interest of the researcher is to find the relationship among the various attributes and their values to determine the termination/continuity status of a patient.

1.4.1.3 Data Preparation and Preprocessing

In the data preparation phase, the final dataset is constructed from the initial raw data. Data preparation tasks were performed in iterative ways. Tasks include description of data sources, carrying out statistical summary measures, finding out distinct values, missing values, outliers, noisy data values. Making decisions on how to handle missing values, outliers and noisy data and data transformation/ reduction activities are also undertaken in this phase, moreover feature/attribute construction or deriving new attributes through segmentation was also made in this phase.

1.4.1.4 Model Building

The data mining models appropriate for the problem under consideration are selected to be classification and Association. For classification purpose, decision tree classifier (J48) in weka for its visualization power of the tree structure and simplicity of understanding is used to model the preprocessed dataset. For Association purpose, Apriori Algorithm in weka was used. Classification model was built after the data is observed for its appropriateness which enabled the researcher to observe how good the dataset is for training and testing purpose. If the size of dataset is sufficient, it is appropriate to be splited to training and testing in which case full training validation method will be used, if not using (k-fold cross validation) for allocating training and testing dataset was the issue which needed decision. A total of 21 experimental model scenarios were built. Eight of them were experimented for classification model and thirteen for association.

1. 4.1.5 Analysis and Evaluation of the models

Sufficiency of the Dataset is evaluated by using Learning Curve by varying the training and testing data used to build the model. Evaluation of the performance of the classifiers is also made in terms of different Confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNP) and False Negative Rate (FNR)), number

correctly classified instances, number of leaves, and size of the trees, execution time and ROC area. The confusion matrix of the classifier models was analyzed in terms of the following variables in Table 1.2.

The formula to compute each of the detailed accuracy measures can be written as:
TPR (Sensitivity) for TRUE Class = TT/TTA. TPR (Specificity) for FALSE Class= FF/TFA. FPR for TRUE Class = FT/TTA = 1- TPR for FALSE Class since TPR for FALSE Class plus FPR for TRUE Class is equal to 1 which means total falsely/wrongly classified. FPR for FALSE Class = 1- TPR for TRUE Class since FPR for FALSE Class plus TPR for TRUE Class is equal to 1. Recall for TRUE Class = TT/TTA. Recall for FALSE Class= FF/TFA. Precision for TRUE Class = TT/TTP. Precision for False Class = TF/TFP.

Table 1.2: Variables Representing Confusion Matrix of the Models

		Predicted		Total
		TRUE	FALSE	
Actual	TRUE	**TT**	**TF**	TTA
	FALSE	**FT**	**FF**	TFA
Total		TTP	TFP	GT

Where: TT (True, True): TP (True Positive) for the True class of the classifier. TF (True, False): FN (False Negative) or False Positive for False Class of the classifier. FT (False, True): FP (False Positive) for the True class of the classifier. FF (False, False): TN (True Negative) or TP for the False class of the classifier. TTP (Total True Predicted = TT + FT). TFP (Total False Predicted = TF + FF). TTA (Total True Actual = TT + TF). TFA (Total Negatives Actual = FF + FT). GT (Grant Total = TTP + TFP or TTA + TFA)

Association rules were evaluated in terms of the number of rules and meaning of patterns generated at different minimum support and confidence thresholds for measuring interestingness of the rules. Association was analyzed in terms of different criteria. The criteria include the number of rules generated at different minimum support and confidence thresholds. The minimum support and confidence thresholds varied from 0.3 and 50% to 0.1 and 100% respectively

1.4.2. Tools

The following tools are used in the research processes:

- **Weka is** used for building models, evaluation and analysis of the models due to availability of the implementation algorithms of model selected in it and prior knowledge of the researcher on it.
- **Ms- Excel** is used for Data preparation, pre-processing and analysis tasks for its filtering capability of attributes with different values
- **Epi Info :** is used for Pre-processing tasks specially for detecting outliers by using box plot and calculation of Interquartile Ranges and other statistical summary measures due to its visualization and analysis power of health related data.
- **Ms-Word** is used for documentation purposes
- **Ms- PowerPoint** is used for preparing slides and presentation purpose

1.4.3 Literature Review

Literature review is conducted to assess data mining technology (both concepts and techniques) and research work in this field. Various books, journals, and articles are consulted to understand the potential applicability of data mining techniques on Health Care data in general and ART care services data in particular.

1.4.4. Ethical Considerations

- The research has nothing to do with the personal identifiers (like name and Address) of individual about whom the data is collected and hence there is no problem of privacy and the confidentiality of individuals.
- The research is purely dedicated to academic purpose (Masters Thesis for the partial fulfillment of M.Sc. degree in Health Informatics).
- Ethical clearance was also obtained from Joint Academic Committee of faculty Medicine and Faculty of Informatics;
- The research is purely for public benefit in general and HIV/ AIDS patients

under study to improve their health in particular;

- The research will not harm anybody in way any
- The research is free of conflict of interest since the researcher has no misleading external pressures like money funding.

1.4.5. Dissemination of Results

Dissemination of the study result will be made through the following mechanisms

- Presentation for the academia (school of public Health and faculty of Informatics) departments
- Publishing in the reputable journals as an article so that it is accessible as additional knowledge of health informatics in the country
- Presentation on different conferences/workshops so that awareness regarding factors affecting ART care follow up (use) will be scaled up;
- Putting the hardcopy in the libraries of concerned organizations so that interested readers can get access to the research output to be used for decision support, take action or to use it as a base for further research in the area;

1.5 Scope and Limitation of the study

Shortage of reference materials in the area of research is main limitation of this research. Anther limitation is generalisablity since only two hospitals were taken due to resource constraints. Moreover, delay of approval of the proposal consumed much of the time that could have been used for actual research work. Lastly the tedious task in this research was cleaning noisy values integration of the records. Lack of standardization in the data entry into the database in the hospitals was serious problem. Data cleaning and integration consumed three-fourth of the research time.

1.6 Organization of the Thesis

This thesis is divided into seven chapters. The first chapter is an introductory part, which contains background to the research work, statement of the problem, objective of the research, and methodology adopted for the study.

The second chapter deals with literature review about data mining technology, methods/techniques used, and its application in the health care sector and related works.

The third chapter is devoted to provide discussions about the different data mining steps that were undertaken in this research work. This includes data preparation activities and model selection, model implementation and descriptions of Experimentations.

The forth chapter deals with model building using different scenarios for both classification and association problems.

Chapter five is devoted to experiment and analysis of classification models using various criteria.

Chapter six is devoted to the experiments and analysis of association

The last (chapter seven) deals with the final concluding remarks and recommendations forwarded based on the research findings.

CHAPTER TWO

DATA MINING TECHNOLOGY AND ITS APPLICATION IN HEALTH CARE

2.1 What is Data Mining?

Concepts and techniques and origins of data mining and its application in the health care are reviewed in this chapter. The review is aimed at providing in depth background about the models to be built to address the research problem.

The need to maximize the use of data for planning and strategic business development in all aspects of management has led many corporations to build comprehensive information systems that record all kind of operational transactions. As these databases grow larger, with gigabytes sizes becoming quite common, they are overwhelming the traditional query and report-based methods of data analysis. On the other hand, data mining is the data driven extraction of information from large databases, a process of automated presentation of patterns, rules, and functions to a knowledgeable users (Piatetsky-Shapiro et. al. 1991). Data mining is becoming increasingly common in the private and public sectors, industries such as banking, insurance, health care, and retailing commonly use data mining to reduce costs, enhance research, and increase efficiency (Seifert J. 2004).

Berry & Linoff, (2000) defined data mining as “a data-driven, exploratory process of knowledge discovery where the focus is on finding and extracting useful patterns of information from large, complex databases”. There are variations in the name- data mining from literature to literature. Hence it is important to get familiar to those variations to have the overall picture of data mining. Fayyad et al. (1996) explained the variation as “Finding useful patterns or meaning in raw data has variously been called KDD (Knowledge Discovery in Databases), data mining, knowledge discovery, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing”. According to Fayyad, Data Mining is “non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.” For this particular research the name data mining and the definition of Fayyad 1996 is used through out this paper.

Data mining has got different functionalities/problems which can be applied on different data. The functionalities include classification, Association, and Clustering and pattern analysis. Seifert J. (2004) described data mining functionalities (problems) and areas of application as follows:

> *Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns), clustering (finding and visually documenting groups of previously unknown facts), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities.*

Data mining, using it roots in statistics, artificial intelligence, databases and visualization techniques, helps to find different kinds of structures and relations in the data, as well as to derive rules and models that enable prediction and decision making in new situations (Han and Kamber, 2006). It is possible to perform classification, estimation, prediction, clustering and description and visualization. As a consequence, data mining has become an essential part of planning the companies' strategies, as a way to increase their profits, retain and gain their markets as well as assessing possible risks. It is used for segmentation of customers; product design improvement; understanding and prediction the customers' preferences identification of risks specifically in health and insurance. Data mining can be applied in all branches of industries, such as: telecommunications, retail, production, banking, tax evasion, education, and health care management (Bach, M., 2007).

Agrawal et al. (1993) also proposed three basic classes of data mining problems namely classification, Association and clustering for applying in health care. The nature of the data mining problem often suggests a certain class of data mining technique or method to

be an appropriate solution. Among the functionalities mentioned above, the researcher's interest is to review literatures related the three main functionalities classification, association, and clustering particularly classification data mining functionality is discussed thoroughly bellow.

2.1.1 Data Mining Classification

Classification predicts categorical (discrete, unordered) labels for records whose class label is unknown. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

Data mining classification is a two-step process: Learning (training) and classification steps. In the training/learning step, a classifier is built describing a predetermined set of data classes or concepts. The classification algorithm builds the classifier model by analyzing or "learning from" training set made up of database records and their associated class labels in training/learning step. A record, X, is represented by an *n*-dimensional attribute (feature) vector, X = (x_1, x_2, … , x_n), depicting *n* measurements made on the records from *n* database attributes, respectively, A_1, A_2, … , A_n. Each record, X, is assumed to belong to a predefined class as determined by another attribute called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical in that each value serves as a category or class. The individual records making up the training set are referred to as training records and are selected from the database under analysis. In the context of classification, data records can be referred to as samples, examples, instances, data points, or objects. Because the class label of each training records is provided, this step is also known as supervised learning (i.e., the learning of the classifier is "supervised" in that it is told to which class each training records belongs). In the second step the model is used for classification. The dataset is split into training and testing data or it will follow iterative process through k-fold cross validation to evaluate accuracy of the classifier. Therefore, a test set is used, made up of test records and their associated class labels. These records are randomly selected from the general data set. They are independent of the training records, meaning that they are not used to construct

the classifier. The accuracy of a classifier on a given test set is the percentage of test set records that are correctly classified by the classifier. The associated class label of each test records is compared with the learned classifier's class prediction for those records (Han, J. and Kamber, M., 2006).

For this research purpose full training set at 50% split for training dataset and the remaining 50% used for testing purpose. At this point the performance is similar to 10-cross validation of weka J48 classifier but the time taken for the 10 fold cross validation is greater than the full training set.

If the accuracy measure is found to be applicable, then the model will be used to predict future class labels at outset. In this research for example, it is possible to predict those who most likely to terminate ART care based on the past pattern from the training model.

Data mining classification problem involves the need to find rules that can partition the data into disjoint groups. Often classification involves supervised data mining tools in which the user is heavily involved in the definition of the different groups and the specification of the rules that can be used to determine to which group a data item belongs. Examples of such tools include decision trees and rule-based techniques. For health care professionals, this type of data mining technique would be important in diagnostic and treatment assistance decision making (Houston, A. et al., 2000)

Classification can be implementations through different classifiers such as ID3, J48, Bayesian belief network and naïve Bayes. In this particular research, decision tree implementation J48 available in weka is selected to be used for classification purpose for the nature of data used for the research can best manipulated with this classifier.

A Decision tree based classier is a convenient way to break large data sets into smaller ones. By presenting a learning set to the root and asking questions at each interior node, the data at the leaves can often be analyzed very simply. After a series of such questions, each leaf can be labeled as one of the class values by using a simple majority vote. Tree based classifiers were independently invented in information theory, statistics, pattern

recognition and machine learning (Grossman, R. 1998)

Decision trees are widely used as a means of generating classification rules because of the existence of a simple but very powerful algorithm called Top-Down Induction of Decision Trees (TDIDT) known well since mid-1960s and has formed the basis for many classification systems, two of the best-known being ID3 and C4.5, as well as being used in many commercial data mining packages (Bramer, M. 2007).

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology (Han, J. and Kamber, M., 2006). Bayesian classifiers are statistical classifiers. Bayesian classifiers include naïve Bayesian and Bayesian belief network. They can predict class membership probabilities, such as the probability that a given record belongs to a particular class. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve”. Naïve bayes classifier is based on the Bayes Theorem, which states “Let X be a data tuple. In Bayesian terms, X is considered “evidence.” As usual, it is described by measurements made on a set of n attributes. Let H be some hypothesis, such that the data tuple X belongs to a specified class C. For classification problems, we want to determine $P(H_j/X)$, the probability that the hypothesis H holds given the “evidence” or observed data tuple X. In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X” (Han, J. and Kamber, M., 2006)

2.1.2 Association Rules

The association data mining problem involves finding all of the rules (or at least a critical subset of rules) for which a particular data attribute is either a consequence (outcome) or an antecedent (precursor). This type of problem is very common in health care professionals who are looking for relationships between diseases and life-styles or demographics or between survival rates and treatments, for example. Association problems are rule-based methods and have confidence factors associated with each rule. Often association-type data mining techniques are employed to help strengthen arguments concerning whether or not to include or eliminate candidate rules from a knowledge model (Houston, A. et al., 2000).

Houston, A. et al., (2000) also stated that association data mining problem involves ordered data, most commonly referred as temporal data. For health care professionals, disease progression and treatment success are two examples of medical information problems where association rule data mining algorithms could be useful. The goal of using associations is to find common relationship amongst the items, or variables, existing in a collection of records.

Let L= $\{x_1 \ldots x_n\}$ be a set of distinct literals, called itemsets. An itemset $X \subseteq L$ with k= $|X|$ is called a k-itemset or simply an itemset. Let a database D be multi-set subsets of L. Each T € D is called a transaction. We say that a transaction T € D supports an itemset $X \subseteq L$ if $X \subseteq T$ holds. An association rule is an expression X => *Y*, where X, Y are itemsets and $X \cap Y = \emptyset$ holds. The fraction of transactions T supporting an itemset X with respect to database D is called the support of X, supp(X) = T € D $|\{X \subseteq T\}| / |D|$. The support of a Rule X=>*Y* is defined as supp (X=>*Y*) = $p(X \cup Y)$ (Agrawal, R. 1994).

The main challenge when mining association rules is the immense number of rules that theoretically must be considered. The number of rules grows exponentially with $|L|$. Since it is neither practical nor desirable to mine such a huge set of rules, the rule sets are typically restricted by minimal thresholds for the quality measures support and confidence, minimum support and minimum confidence respectively. The support

(prevalence) of a rule is the proportion of observations that contain the item set of the rule. The confidence is the conditional probability of itemset B given itemset A, P (Item B/ Item A). Levin and Zahavi (1999) argued that a rule is "interesting" if the conditional probability P (Itemset B/ Itemset A) is significantly different than P (Itemset B). Support and Confidence are two probability measures, which are used to assess the associations of itemsets.

2.1.3 Clustering

Clustering is a data mining technique that separates dataset into groups whose members belong together. Each object is assigned to the group it is most similar to. This is similar to assigning animals and plants into families where the members are alike. Clustering does not require a prior knowledge of the groups that are formed and the members who must belong to it. It is technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering (Han, J. and Kamber, M., 2006). Han and Kamber explained K-Means algorithm as one of the most common clustering algorithms that groups data with similar characteristics or features together. These groups of data are called clusters. The data in a cluster will have similar features or characteristics which is dissimilar from the data in other clusters.

2.1.4 Origins of Data Mining

Recently data mining has been the subject of much research in business and software magazines. However, just a few years ago, only few people had heard of the term data mining. Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s. This section explores the history of data mining.

Data mining roots are traced back along three family lines. The longest of these three lines is classical statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Classical statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role (Kamber and Han, 2006)

Data mining's second longest family line is Artificial Intelligence (AI). This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. The notable exceptions were certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS) (Bramer M., 2007).

The third family line of data mining is machine learning, which is more accurately described as the union of statistics and AI. While AI was not a commercial success, its techniques were largely co-opted by machine learning. Machine learning, able to take advantage of the ever-improving price/performance ratios offered by computers of the 80s and 90s, found more applications because the entry price was lower than AI. Machine learning could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals (Agrawal, 1993).

Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within it. Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to discover trends which they could not otherwise find. As it is mentioned above data mining mingles different aspects from different disciplines to extract non trivial and hidden knowledge out of the data stored for period of time on certain subject (Bach, M., 2007).

2.1.5 Data Mining Processes

Discovering knowledge from data should be seen as a process containing several steps like understanding the domain, preparing the data set, discovering patterns (data mining), post processing of discovered patterns, and putting the results into use (Mannila , 2002).

(IRM, 1999) also stated that data mining involves a number of steps: setting goals, data selection, data preparation and preprocessing, model building and analysis, validation, predictions and presentation.

(IRM, 1999) keeps on explaining each step as follows:

> *At the beginning, it is important to understand what the goals of the data mining exercise are. Data selection involves careful isolation of variables from one or more databases. It is better to start with less data than more, and to evolve a more complex model from a successful simple one. Once the raw data has been selected, preprocessing is often necessary. Missing or extreme (outliers) values must be identified. If the data is not numeric, coding schemes must be employed before the modified data can be used by the technical tools, such as cluster analysis. Analysis proceeds using one or more of the tools available for data mining.*

As in many problem-solving researches, the data-mining research starts with a clear definition of the business problem involved and the objective function. Once the business

problem and the objective of the data mining research are clearly defined, the next step is to select the target dataset for analysis. This requires figuring out what data are needed, which data are most important and integrating the information. One needs to extract the target data to analyze in a way that is consistent with the business problem involved and the objective of the research. Data Preparation and preprocessing of the selected dataset is the next step, which is often the most time consuming task of the data mining process, especially if data is drawn directly from the company's operational database rather than from data warehouse (Levin & Zahavi, 1999). The result of data preprocessing is a dataset, which is ready to build the data mining model.

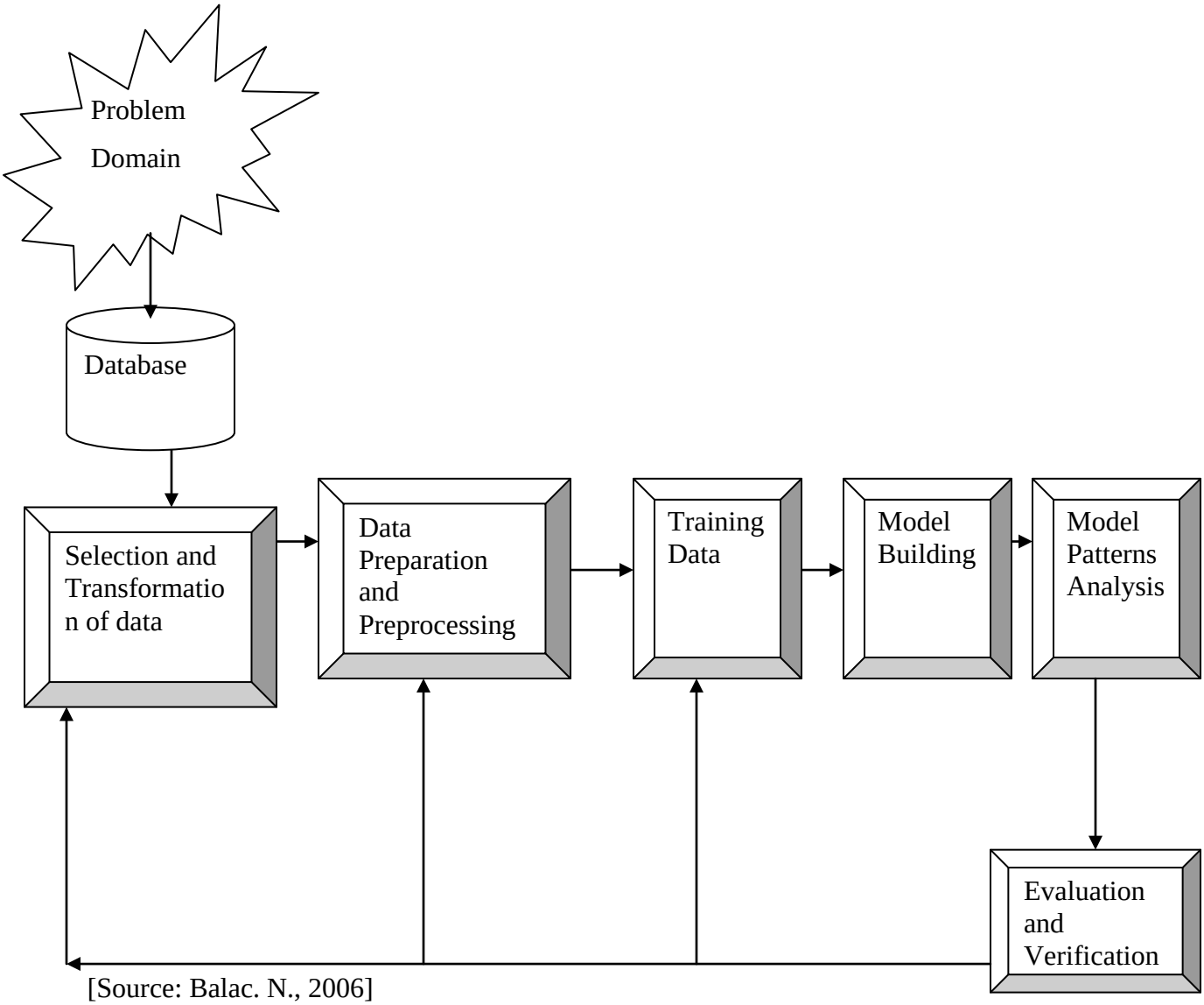
The data-mining modeling is the next step once we get the preprocessed dataset. It is only at this point that one invokes data mining models and tools to interrogate the data and convert it into knowledge for decision making. This model-building step involves selecting data mining tools, transforming the data if the tool requires it, generating samples for training and testing the model, and finally using the tools to build and select a model. The resulting models might have important patterns to be analyzed and interpreted to be used as decision support knowledge (Kamber and Han, 2006).

Multiple techniques are used, and the analysis step is an iterative and experimental one. Once analysis has found new patterns, validation is necessary to confirm these patterns can be profitably exploited. Finally, presentation of the results is essential to show management both the results of the data mining and success of its predictions. Data visualization techniques are vital to this step. Pictures, not words, are often the bottom line for data mining (IRM, 1999).

Therefore it possible to understand from the above discussion that data mining is complicated process that has to be planned very carefully in order to be successful. Planning involves identification of a research problem, data selection, data preparation and preprocessing, selection and transformation of variables, modeling, validation of the model, and model analysis and application.

Data mining processes starts from selecting the domain area (problem area) on which the research is conducted. Following the identification of the research area and data sources

(database); data selected may be transformed according to nature of data mining tool we have at hand. Data preparation and preprocessing is needed for eliminating noisy, missing and outlier values from which training data is obtained. Data mining tool is applied on the preprocessed dataset which results in the models whose pattern is analyzed and interpreted by the researcher. Based on evaluation/verification of the discovered knowledge; the result of model either deployed or revisiting the previous steps for possible corrective actions (Balac, N., 2006). For clarity it is possible to look at the following illustration of data mining processes.



[Source: Balac. N., 2006]

Fig. 2.1: Data Mining Processes

CRISP- DM which is a methodology adapted for this research purpose has also got certain steps. The life cycle of a data mining research consists of six phases according CRISP-DM. The sequence of the phases is not rigid. Moving back and forth between different phases as necessary is required. According to CRISP-DM, data mining is not a solution at one go. Rather, the lessons learned during the process and the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes benefit from the experiences of previous ones (Chapman, P. et. al., 2000)

The phases of the CRISP-DM are depicted in the following illustration.

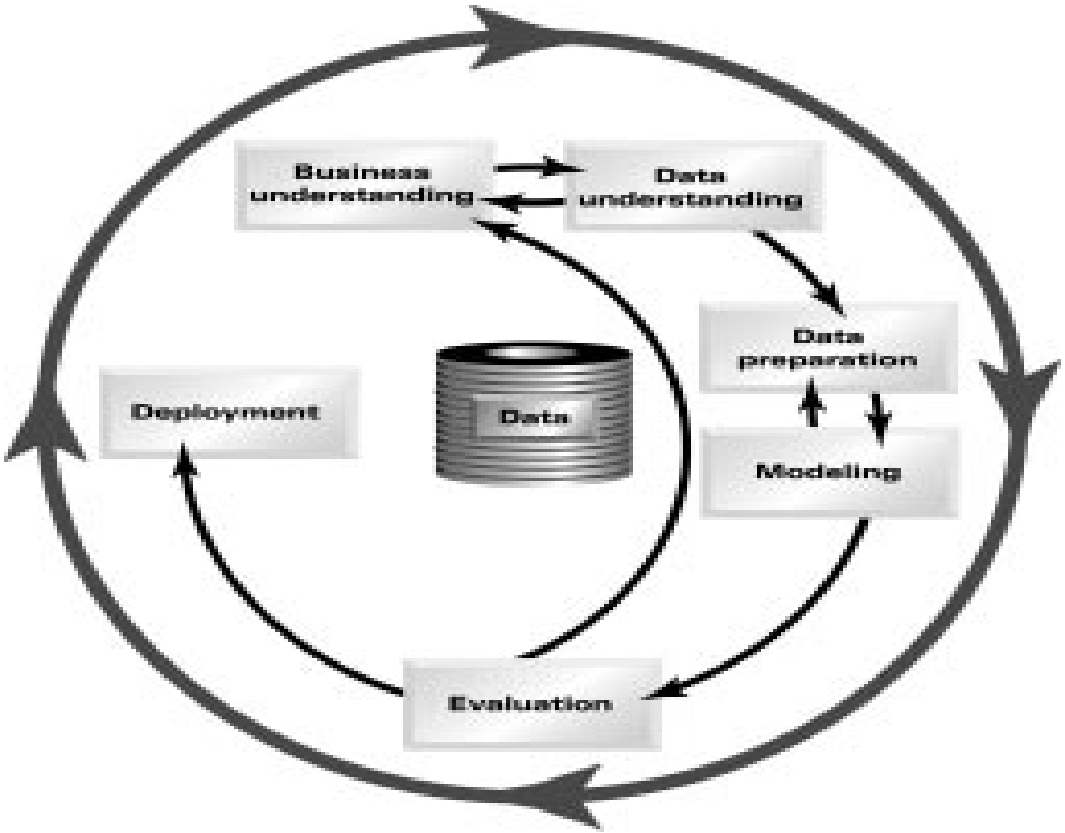


Fig. 2.2: Phases of CRISP-DM

[Source: Chapman, P. et. al., (2000)]

2.2 Application of Data Mining in Health Care

Extensive amounts of data gathered in heath care databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. Medical informatics may use the technologies developed in the new interdisciplinary field of knowledge discovery in databases (KDD) which is referred as data mining in this research. Practical use of knowledge management technologies can contribute a lot to decision support systems in health care. Data mining technology provides a user- oriented approach to novel and hidden patterns in the health care data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service, and by the medical practitioners to reduce the number of adverse drug effect and to suggest less expensive therapeutically equivalent alternatives. Hence it possible to say data mining techniques can be applied to create knowledge rich health care environment (Kaur, H. and Wasan S. 2006)

According to Kaur H. and Wasan S. (2006) areas in which data mining techniques are applicable in health care include: Executive Information System (EIS) for health care, Forecasting treatment costs and demand of resources, anticipating patient's future behavior; given their history, Public Health Informatics (PHI), E-governance structures in health care, risk identification for certain disease and Health Insurance.

The benefits of data mining and precautions that should be taken in making decisions in health care setting is stated by Baylis P. (2006) as follows:

> *The use of data mining has focused on evidence-based patterns from previous patient treatment. In all likelihood, the absence of automated discovery of patterns would leave many questions unasked. These questions, if asked, would benefit not only the resource utilization for patient treatment, but also the health of the patient. Data mining helps professionals discover these patterns and put them to work. As models are based directly on history of patients, they represent the ultimate in evidence-based care.*

Today, numerous health care organizations are using data mining tools and techniques in order to raise the quality and efficiency of health-related products and services. A number of articles published in the health care literature revealed the practical application of data mining techniques in the analysis of health information in developed world.

In an effort to turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficiency of patient care. Data mining can be used to help predict future patient behavior and to improve treatment programs. By identifying high-risk patients, clinicians can better manage the care of patients today so they do not become the problems of tomorrow (Rogers and Joyner, 2002).

Regarding ART services and associated concerns, Morris, J. et. al. (2007) stated as follows:

> *Antiretroviral Therapy (ART) has been unequivocally associated with improved survival among individuals infected with human immunodeficiency virus (HIV) and has decreased the incidence of AIDS-associated opportunistic infections. However, as HIV-infected patients begin to live longer on ART therapy, issues regarding viral resistance, short and long-term drug toxicities, and adherence due to complex regimens and other demographic and behavior issues have become important concerns. These issues, in addition to the increasing number of treatment options, make it often necessary to modify therapy in order to achieve the goals of viral suppression, longer survival, and improved quality of life. These modifications in treatment, either switching to a new regimen consisting of a different drug or drugs, downshifting to a non-HAART treatment regimen, or discontinuation of antiretroviral use, have uncertain effects on the entire course of disease and may limit the number of future treatment options*

.

Hence applying data mining techniques will help to find hidden patterns and relationship that exist with in the data collected during the provision of ART care services. This fact is stated by Vararuk et. al. (2008) as follows:

The information explosion has created a situation where health care providers are unable to assimilate the volumes of information available to make decisions in optimal fashion. In addition to information explosion, the health care provider lacks guidance for HIV/AIDS assessment that has been scientific and systematic other than traditional statistical methods while the information needed is a nontrivial research problem that is overwhelmed by incomplete clinical data. Hence the need for the application of data mining techniques to come up with nontrivial and hidden knowledge or patterns out of large volume of Clinical dataset becomes important since data mining techniques offer improved decision support for HIV/AIDS, and the opportunistic infections (symptoms)

2.2.1 Related Works

Vararuk et. al. (2008) investigated patterns in HIV/AIDS patient data through the use of data mining techniques in Thailand. As the researchers stated, patterns obtained in their study can be used for better management of the disease and more appropriate targeting of resources. They took a total of 250,000 anonymised records from HIV/AIDS patients imported into a database. IBM's Intelligent Miner was used for clustering and association rule discovery. They stated that clustering highlighted groups of patients with common characteristics. Unexpected association rules were identified that were not expected in the data and were different from traditional reporting mechanisms utilized by medical practitioners. It also allowed the identification of symptoms that co-exist or are precursors of other symptoms.

The significance of the research was stated as identification of symptoms that are precursors of other symptoms can allow the targeting of the former so that the later symptoms can be avoided. Their study also showed that providing a realistic and targeted approach to the management of resources available for HIV/AIDS treatment can provide a much better service, while at the same time reducing the expense of that service.

Bach, M. and Ćosić D. (2007) conducted a research on benefits of data mining in health care management. The aim of the study was to show the benefits of data mining in health care management in general and tried to show a way to raise awareness of women in terms of contraceptive methods they use (do not use) in particular.

Methods used for data mining analysis to determine if there are common characteristics of the women according to their choice of contraception, which is classification problem. They used decision trees classifier for classification purpose.

Bach and Ćosić modeled their data using a tool called Statistica, based on the database that was formed as a result of an Indonesian research that was conducted in 1987. The sample they used contains married women who were either not pregnant or did not know if they were pregnant at the time of the interview. The database consists of 1473 cases. Extensive internet search was conducted in order to detect a number of articles cited in scientific databases published on the subject of data mining in health care management.

Results they obtained showed that the most important variable in case of Women's choice of contraceptive methods is – a husband's profession. They also retrieved 221 articles published on the application of data mining in health care.

In their conclusion, they stated that the goal of their study is achieved in two ways: first, retrieving 221 articles published on the subject they have proved the benefits of data mining in the health care management. Second, the decision tree method is successfully applied in explanation of women's choice of contraceptive methods.

Kaur,H. and Wasan, S (2000) examine the potential use of classification based data mining techniques such as rule based, decision tree to massive volume of healthcare data. In particular they considered a case study using classification techniques on a medical dataset of diabetic patients. They concluded that data mining classification offered very important signal for the diabetics care services providers as given patient records with corresponding diagnosis, knowledge discovered through data mining methods are enabled the health care worker to diagnose new cases.

Morris J. et. al. (2007) conducted a statistical research with a title “Injection Drug Use and Patterns of Highly Active Antiretroviral Therapy (HAART) use: An analysis of ALIVE, WIHS, and MACS cohorts” in different states of US. In their introduction they explained that Sustained use of antiretroviral therapy has been consistently shown to be one of the primary predictors of long-term effectiveness. Switching and discontinuation reflect patient and provider decisions that may limit future treatment options. In their study, they utilized data reported at semi-annual study visits from three prospective cohort studies, the AIDS Link to IntraVenous Exposure (ALIVE), the Women's Interagency HIV Study (WIHS), and the Multicenter AIDS Cohort Study (MACS), to investigate determinants of HAART modification with a particular focus on reported injection drug use (IDU)

In their method they stated that they used longitudinal data collected between 1996 and 2004 contributed from 2,266 participants (37% with a reported history of IDU) who reported initiating their first HAART regimen during follow-up. Separate proportional-hazards models were used to identify factors measured prior to HAART-initiation associated with the time to first HAART discontinuation and first switch of components of HAART among continuous HAART users.

Shagaw Anagew (2002) has conducted data mining research in health care with the title “Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project (BRHP)” with aim to investigate the potential applicability of data mining technology in developing a model that can support primary health care providers, policy makers, planners. He applied Neural Network and Decision tree Classifiers for classification purpose using BrainMaker tools. He followed three steps in conducting the research. These were collecting of data, data preparation and model building and testing. The required data was selected and extracted from the ten years surveillance dataset of the BRHP epidemiological study.

Shegaw stated that several neural networks and decision tree models were built and tested for their classification accuracy and many models with encouraging results were obtained. Models were built and tested by using a sample dataset of 1100 records of both

alive and died children. Shegaw stated also that the two data mining methods used for the research work have proved to yield comparably sufficient results for practical use as far as misclassification rates come into consideration.

Shegaw compared the two methods used as “unlike the neural network models, the results obtained by using the decision tree approach provided simple rules that can be used by non technical health care professionals to identify cases for which the rule is applicable”. Finally, he proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities.

Abraham Tesso (2005) has also conducted data mining research in health care in general and on HIV/AIDS in particular with the title- “Application of Data Mining Technology to identify determinant factors of HIV Infection and to find their Association Rule”: Case of Center for Disease Control and Prevention (CDC) with the objective to see the potential applicability of data mining on VCT data to broaden the insight regarding determinant factors of HIV/AIDS infection. Abrham used 5267 records of VCT service visitors for the research.

Abreham has gone through various data preprocessing activities to come up with dataset used to build the model using association rule by Apriori Algorithms. He used 70% of the total number of records for training purpose and the remaining 30% for Validation purpose on the unseen data. KnowledgeSTUDIO and Weka were used by Abreham for experimenting the research. Abrham reported that his research output revealed promising result, which will be useable by health professionals, government, policy makers and the society at large.

CHAPTER THREE

DATA PREPROCESSING AND MODEL SELECTION

Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. There are many possible reasons for noisy data (having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Therefore; the data preparation and preprocessing to clean noisy data, to handle missing values and inconsistent attribute values has an immense importance in data mining research (Han and Kamber, 2006).

Therefore; this chapter is devoted to describe the data sources, how the data is taken from the source databases, different statistical summary measures to see what is missing, what inconsistencies exist in the data, outliers and related decisions to curb the data quality problems**.**

3.1 Data Source Description

Data for this research is collected from Adama and Asella Hospitals ART databases. The databases of both hospitals have the same format. The database is on SQL server Database Management System (DBMS). The databases have the interface through which the data clerks can enter patient data and generate different reports.

There were 13486 records in the database of Adama Hospital, collected from 1993 to 2001 E.C inclusive. There were also 5254 records in the database of Asella Hospitals. The total number of records is 18740. All the records are taken from these databases

starting from 1993 to 2001 E.C, since data mining research can be conducted even huge number of records more than this. The summary of the data sources is as illustrated in the Table 3.1 below:

Table 3.1: Source and Number of Records

Hospital	Frequency	Percent
Adama	13486	72
Asella	5254	28
Total	18740	100
Distinct values	2	
Type	Nominal	

There are two files used from the database for the research purpose: All Patient file and patient list file.

The attributes in the patient list file include Item Number, Card Number, Registration Date, Eligible Date, Eligible Ready Date, Age, Sex, ART Stage, Functional Status, On ART Date and Termination status**.** More description of these attributes will be given later in this section on Table 3.2.

From all patient detail file**,** Marital Status, Educational Level, Religion, Occupation, WHO stage ('OAWHO') and CD4 count ('OACD4') are joined to the patient list by card number to have complete record of individual patient as required for the problem as explained in the methodology section. The full description of the attributes taken from the database is given in Table 3.2.

Table 3.2: Description of the Attributes

S.No	Attributes	Meaning	Values	Data type
1	Hospital*	Hospital from which the Data is taken	Adama Asella	Nominal
2	Item No	Sequential number of the patient	Integer values starting from 1	Numeric
3	Card No	Card Number of the patient	Numeric value and year in which the patient is registered separated by the forward slash	Nominal
4	Registration Date	Date on which the patient is registered for the service	Data in format dd-mm-yy in E.C	Date
5	Age	Age of the patient	Numeric age value	Numeric
6	Sex	Sex of the patient	F- Female M- Male	Nominal
7	ART Stage	Stage of ART care	OA –On taking ARV drug IN- In Care for other disease EL- Eligible to be on ARV Drug ER- Eligible and Ready to start ARV Drug	Nominal
8	Functional Status	Status of the Patient while starting the care	A-Ambulatory (perform activities of daily living) W-Working (able to perform usual work in or out of the house ,harvest, go to school , plying) B-Bedridden(Not perform activities of daily live) P-Appropriate(satisfies age milestones) D- Delay(failure to attain millstone for age) R- Regression(Loss of what has been attained for age)	Nominal

S.No	Attributes	Meaning	Values	Data type
9	**EL Date**	Eligible Date	Date in form DD-MM-YY	Date
10	**ER Date**	Eligible and Ready Date	Date in form DD-MM-YY	Date
11	**OA Date**	Date on which the patients is on ARV	Date in form DD-MM-YY	Date
12	**Marital Status**	Marital Status of the patient	✓ Never Married ✓ Married ✓ Living together ✓ Divorced ✓ Separated ✓ Widower ✓ Others	Nominal
13	**Educational Level**	Educational Level of the patient	✓ No Education ✓ Primary ✓ Secondary ✓ Tertiary ✓ Other	Nominal
14	**Religion**	Religion of the patient	✓ Orthodox ✓ Muslim ✓ Protestant ✓ Catholic ✓ Other	Nominal
15	**Occupation**	Occupation of the patient	1.Employed 2.Self Employed 3. Unemployed 4. Student 5.Other	Nominal
16	**AOWHO**	WHO Stage at which the patient is on ART	1. Stage 1 2. Stage 2 3. Stage 3 4. Stage 4	Numeric
17	**OACD4**	CD4 count values of the patient used as a base line to on ART	None Zero positive Number	Numeric
18	**Termination**	Termination Status of the patients on ART	False-On ART True- Not on ART	Nominal
19	**Year ***	Year in which the patient is put on its respective stage	Year in E. C	Numeric
20	**Mon-Year***	Month	mm-yyyy	Date

* The ‘Year’ and ‘Mon-Year’ attributes were derived from the registration date of the patient. Another newly generated attribute is the ‘Hospital’ attribute, which has got two values (Adama and Asella). All the attributes that is obtained through the search have been taken and described as in Table 3.2.

3.2 Descriptive Statistical Summary of Attributes

3.2.1 ARTStage

As it has been described above in the attributes description table, the ART Stage stands for ART stage and it is nominal type attribute. This attribute has 4 valid attribute values. These are EL (Eligible), ER (Eligible and Ready), IN (In Care), OA (On ART). The frequency table of the attribute looks like as in the Table 3.3.

Table 3.3: Statistical Summary of ART Stage Attribute

ARTStage: Nominal			
		Frequency	Percent
Distinct values	EL	416	2
	ER	25	0
	IN	7037	38
	OA	11262	60
	Total	18740	100

As you can see from table 3.3, the modal value for this attribute is OA. Different literatures recommend that the missing vales for the nominal type attribute shall be replaced by the modal value. This attribute has 4 valid attribute values, which is to mean the distinct values.

3.2.2 Marital Status

The frequency Table of this attribute and the possible nominal values of the attribute is shown in Table 3.4.

Table 3.4: Statistical summary of Marital Status Attribute

MaritStatus: Nominal			
		Frequency	Percent
Missing		1360	7.26
Distinct Values	Divorced	2179	11.63
	Living Together	2	0.01
	Married	7486	39.95
	Never Married	4256	22.71
	Others	2	0.01
	Separated	773	4.12
	Widower	2682	14.31
Total		18740	100

As you can see from the table above, the modal value for marital status is married. The attribute has seven valid values (Divorced, Living together, married, never married, others, widower and separated). The missing values of the attribute are 1360(7%). Most of the missing values for the attribute are found in those patients records whose age is less than 18 years and it seems logical to be inferred to be Never Married and in this case it can not be appropriate to replace the most frequent value. The researcher decided to replace the missing values with the 'Never Married' value for those whose age is less than 15 years and the modal value (Married) for those whose age is greater than 18.

3.2.3 Religion Attribute

Religion is a nominal attribute having five distinct vales attributes (Catholic, Muslim, Orthodox, Protestant and others). The statistical summary measures of this attribute are in table 3.5 bellow.

Table 3.5: Statistical Summary of Religion Attribute

Religion: Nominal			
		Frequency	Percent
Missing Values		2880	15.37
Distinct values	Catholic	30	0.16
	Muslim	2155	11.50
	Orthodox	12235	65.29
	Other	62	0.33
	Protestant	1378	7.35
Total		18740	100

As it can be seen from table 3.5 above, the modal value is Orthodox and the missing values in this attribute are 2880(15%). The missing values in this attribute can be handled by replacing with the most frequent value (mode) which is 'Orthodox'.

3.2.4 Sex Attribute

Sex is also nominal attribute of the patients, having two valid values Female and Male. As we can see from table 6 bellow, the modal value is 'Female', which is 57%. The missing values in this attribute are only 7. It can be replaced by the most frequent value of the attribute, which is 'Female'. The statistics is presented as in Table 3.6.

Table 3.6: Statistical Summary of Sex Attribute

Sex			
		Frequency	Percent
Missing		7	0.04
Distinct	F	10705	57.12
	M	8028	42.84
Total		18740	100

3.2.5. Educational Level

Educational Level is nominal attribute having fife valid values (No Education, Primary, Secondary, Tertiary and other). The most frequent value for Educational Level of the patient is 'Primary' which is 32%. The missing values for this attribute are 2883(15%). Since it is nominal attribute, it is logical to replace the missing values with the most frequent value (Primary). For the clear statistical summary of the attribute, see Table; 3.7.

Table 3.7: Statistical summary of Educational level Attribute of the Patients

Educational Level: Nominal			
		Frequency	Percent
Missing Values		2883	15.38
Distinct	No Education	4260	22.73
	Others	7	0.04
	Primary	5929	31.64
	Secondary	4711	25.14
	Tertiary	950	5.07
Total		18740	100

3.2.6 OAWHO Attribute

OAWHO in this dataset refers to WHO stage when the patient is put on ART service. This attribute has a noisy data which is 0. There is no WHO stage called stage 0. The WHO stage starts at 1 and ranges to 4 as it has already been discussed in the background section. However, the patient is not waiting the stage criteria, and directly allowed to start the care due to serious health problem; the value is missed (0) in the database. Based on the discussion with the domain experts, it is good to consider the 0 values of this attribute as missing value. The percentage of the missing values of the attribute is 7139. The

missing values were decided to replace by the most frequent value, stage 3. See table 3.8 for the statistical summary.

Table 3.8: Statistical summary of OAWHO attribute

OAWHO: Nominal			
		Frequency	Percent
	Missing values	7139	38
Distinct Values	1	798	4
	2	1876	10
	3	7087	38
	4	1840	10
	Total	16989	91
Total		18740	100

3.2.7 Occupation Attribute

It refers to the occupation status of an individual patient who was in the care. It has five valid values (1 (employed), 2 (self employed), 3 (Unemployed), 4 (student) and 5 (others)). The most frequent value is '3' (unemployed) and the missing values for the attribute are 2367 (12%). It has to be replaced by '3' which stands for 'unemployed' value of the attribute. For further statistical summary, look at the Table 3.9

Table 3.9: Statistical Summary Measure for the Occupation Attribute

Occupation: Nominal			
		Frequency	Percent
	Missing	2367	12.63
Distinct Values	1	1445	7.71
	2	1817	9.70
	3	10905	58.19
	4	179	0.96
	5	2027	10.82
Total		16373	87.37
Total		18740	100

3.2 .9 Functional Status Attribute

Functional Status in the dataset refers to health condition of the patient when they are starting the ART care service. It has six valid values as it can be observed from table 3.10. These values have been described in table 3.2 in the attribute description table above. The mode value is ‘W’ which is to mean working. The missing values are 623 (3%) which can be replaced by the most frequent value ‘W’. The statistics is presented in Table 3.10.

Table 3.10: Statistical Summary of Functional Status Attribute

Functional Status: Nominal			
		Frequency	Percent
Missing		623	3.3
Distinct values	A	2922	15.6
	B	1051	5.6
	D	116	0.6
	P	1049	5.6
	R	17	0.1
	W	12962	69.2
	Total	18740	100

3.2.10 Year Attribute

Year attribute is derived attribute from the registration date. It is derived as a conceptual hierarchy of the date so that more general information will be obtained related to the termination status. The year is in Ethiopian calendar. As we can see from Table 3.11, the beginning date obtained is 1993 E.C. the last year is 2001. The mode is 2000 (the year in which many of the patients came to the service) which is 28%.

Table 3.11: Statistical Summary of Year Attribute

Year: Number			
		Frequency	Percent
Distinct values	1993	3	0.02
	1994	2	0.01
	1995	20	0.11
	1996	76	0.41
	1997	577	3.08
	1998	3268	17.44
	1999	4695	25.05
	2000	5261	28.07
	2001	4838	25.82
Total		18740	100

3. 2.11 Termination Attribute

This attribute is nominal in type. Termination attribute refers to the termination status of the patients who started the ART care. The valid values for the attribute are ‘True’ to mean the patient has terminated the care and ‘False’ to mean the patient is on the ART care. As it is already mentioned in the methodology section in chapter one, this attribute is a class label (Dependent Variable). The most frequent value for the attribute is ‘False’ which is 73.4%. Statistical summary of the ‘Termination’ attribute is presented in table 3.12:

Table 3.12: Statistical summary of the Termination Status

		Frequency	Percent
Distinct Values	False	13804	74%
	True	4936	26%
Total		18740	100

3. 2.12 Month-Year

Mon-year is generated attributed which a conceptual hierarchy of the date on which particular patient is registered for the care. It has 89 distinct values. It has no missing values. The frequency table is too long to present in this section. The month-year in which many patients were registered was 11-2000, which accounted for 3.9% of the total population, came to the facilities under investigation.

3.2.13 OACD4 Attribute

OACD4 is a numeric attribute which refers to the CD4 count of the patients who are in the care. It has 564 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 152 as calculated on none 0 (non valid value). The zero values were decided to be replaced by the mean value. The attribute needs descretization as its distinct values are too much. We can see table 3.13 for important statistical summary. The frequency table is to long since we have 571 distinct Values. To detect records with outlier values, fife number summary was done on the none zero values i.e. only for those patients CD4 count was made accordingly; Q1 is 66 and Q3 is 194. The number of records detected as outlier was 244.

Table 3.13: Statistical Summary of OACD4 Attribute

CD4		
Missing		123011
Mean		152
Median		129
Mode		40
Minimum		1
Maximum		2000
Percentiles	25%	66
	50%	129
	75%	194
	IQR	128
	1.5*IQR	192
	Upper limit	386

The box plot is also drawn for OACD4 attributes in effort to find the outlier values for the attribute. The box plot looks like as in the figure 2 bellow. The IQR is 128. The values we have above Q3 + (1.5*IQR) can be considered as outlier values as many literatures recommend. In this case: 194 + 192= 386. Therefore, values for OACD4 above 386 can be considered outlier values. The 244 outlier values for this attribute will be removed and replaced by the mean value (152).

3.2.14 Age Attribute

This attribute is numeric type and holds numeric age values in years of individual patient in the ART care service. Its IQR is 13. The most frequent age is 30 years. The number of missing values for the attribute is 10. The missing values can be replaced by mean age, which is 31 years. The number of values which were found outlier is 106. The researcher's decision on the outlier values to remove and replaced by the mean value (31 years).

Table 3.14: Statistical Summary of Age Attribute

Age: Numeric		
Missing		10
Mean		31.28
Median		30
Mode		30
Std. Deviation		12.112
Variance		146.712
Range		80
Minimum		0
Maximum		80
Percentiles	25	25
	50	30
	75	38

Box plot graph is also plotted to visually depict the outliers. The third Quartile (Q3) is 38, the first quartile is 25. The IQR is 13 as presented in the table above 1.5 * IQR is 20 years. So 38+20 is 58 years is the upper limit for outliers i.e. age values beyond 58 years are outliers. For the lower limit is Q1- (1.5*IQR) which is 25-20 = 5. Age values bellow 5 years can be considered outlier in this dataset.

3.3 Data Cleaning

Data Cleaning refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques, for example) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics). Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning/training (Han and Kamber, 2006). In this subsection, the researcher tries to carryout different data cleaning tasks based on the statistical summary measures discussed above in section 3.2.

3.3.1 Handling Missing Values

As it has been shown above in the statistical summary of the attributes in the dataset, there are missing values in all except 'Hospital' and 'Termination', 'Year' and 'Mon-Year' attributes. In this subsection, the researcher tries to explain how those missing values were handled. There are different mechanisms to handle missing values in data mining techniques.

To handle the problem of missing values for nominal variables, replacing with modal value is recommended in Two Crows Corporation (1999). For the numeric type attribute the missing values were recommended to be replaced by the Mean value.

Therefore, in the dataset collected for this research work, both nominal and numeric attributes missing values were handled in accordance with the above suggestion. The attributes are: 'OAWHO', 'Educational Level', 'Religion', 'Occupation', 'ART Stage', 'Functional Status' and "Sex". Since all the above fields are nominal variables, for any missing values in those fields, the modal (most frequent) value was used. For 'Marital Status' attribute, special consideration was made according to the patients age. For those patients whose age is below 18 years and having missing values in the 'Marital Status' were replaced by 'never married'. Those patients whose age is above 18 years, the missing values for the attribute is replaced by the most frequent value. The modal values for the above nominal are summarized in Table 3.15.

Table 3.15: Summary of Modal values for Nominal Attributes with missing Values

Attributes	Modal Value
OAWHO	Stage 3
Marital Status	Married
Educational Level	Primary
Religion	Orthodox
Sex	Female
ARTStage	OA
Occupation	unemployed

The mean value for age attribute is 31 years and the mean for the OACD4 is 152. Therefore based on the above recommendations, mean value is used for OCD4 attribute to replace the outlier values. For the 'Age' attribute, missing values were handled according to the recommendation made under section 3.2.13.

3.3.2 Handling Outliers Values

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group. The degree to which numerical 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 interquartile 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, 2006).

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

As stated in Han and Kamber (2006); a common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5*IQR above the third quartile 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 *IRQ)

will be considered outlier values. According to this recommendation, Age and OACD4, which are numeric in type, have a kind of outlier as stated under sections 3.2.13 and 3.2.14. The box plot for age attributes was also illustrated in fig. 3.1. The outlier limits for the OACD4 lies between 1 and 386. The lower and upper limit for the outliers in the Age attribute is 5 and 58 respectively. See section 3.2.13 and 3.2.14 for the detail of the calculation to find the outliers.

Decision on the outlier values by the researcher is to remove the outlier values and to replace them with the mean values of the attributes as recommended in (Kamber and Han, 2006).

3.3.3 Handling Noisy Values

Noise is a random error or variance in a measured variable. Noisy values were obtained in Age, OAHWO attributes. While, both attributes have no valid value called 0; it appeared in them. In case of age those children whose age is less than 6 months were assumed 0 year by the database as it was it was obtained from the database clerks. This value was also happened the outlier value and is handled as mentioned above under section 3.3.2.

For 0 values of 'OAWHO' attribute, for those patients who were put on the care without waiting the staging criteria. It was considered as missing values and replaced by the most frequent value which is stage 3.

3.3.4 Data Transformation and Reduction

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve, smoothing or feature (attribute) construction, which works to remove noise from the data. Smoothing techniques include binning, regression, and clustering. Attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process. Smoothing can also serve as data reduction, for example in the case

of smoothing through binning, since the number of the distinct values for a certain attribute is reduced (Han and Kamber, 2006).

In this dataset the ‘Age’ and ‘OACD4’ attributes were descretized (binned) both to reduce the distinct values of the attributes so that it will suit the mining tool and to obtain meaningful patterns. Feature (attribute) construction tasks were also done by the researcher. The attributes ‘Hospital’, ‘Year’ and Mon-Year’ are constructed features in fact; the later two attributes were constructed through conceptual hierarchy generation through segmentation of the registration date. What makes it conceptual hierarchy is the fact that the researcher brought the dates into the months and the years. Since the data is taken from two sources (hospitals), the researcher constructed new attribute called ‘Hospital’ to hold values ‘Adama’ and ‘Asella’, in case we might discover some kind of pattern of termination status by hospital .

The Eligible date, Eligible and Ready date and On ART date are not complete across the entire column. Rather; they are registered for patients on their respective Stages. Therefore; it was found important that to bring into a single column as date for particular stage and the rest of the date columns were removed. Then since there was no difference in registration year and staging year, the researcher decided to remove the staging date attribute and take year and month-year of the registration date. On top of this, we are interested in the registration year as stated in the objective section 1.3 since it tells us the year in which the patient begin the service.

Item No and ‘Card No’ are also related in the sense that they are referring to the order in which the patients started the treatment except we have the year as the last element of the card ‘Card No’. The distinct values both Card No and the Item No were almost as many as the number of records and hence has no contribution to the date mining task. Therefore; both ‘Card No’ and ‘Item No’ were removed from the dataset.

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels

can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to a concise, easy-to-use, knowledge-level representation of mining results. A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as Childhood, adolescent, adult and old) (Han and Kamber, 2006).

Therefore the researcher carried out descretization on the 'age' and 'OACD4' numeric attributes as per recommended under their statistical summary sections using binning methods on weka.

Binning is a top-down splitting technique based on a specified number of bins. Equal-width (distance) partitioning: It divides the range into N intervals of equal size: uniform grid. If A and B are the lowest and highest values of the attribute respectively, the width of intervals will will be calculated as:

$$W = (B-A)/N$$

This approach leads into bins with non uniform distribution of data elements per bin. This binning method is available in weka and applied on these attributes. The following Table 3.16 shows the descretized labels of age attributes

Table 3.16: Discretized result of Age attributes

Labels	Frequencies
5--18	1166
18--32	9037
32--45	6387
45--58	2150
Total	18740

As we can see from the figure above, the 79 distinct values reported in the statistical summary section of the attribute is reduced into four labels through binning using equal width, which revealed the bins 5-18, 18-32, 32-45 and 45-58 with their frequencies as indicated in the table above. The bins decided to be 4 bins for the intention that it will be good age category like Youth, adolescent, Adult, old, which would provide meaningful for interpretation.

The result of the discretization of the OACD4 attribute looks like the following in fig. 3.3. As you can see from the statistical summary section, this attribute had 564 distinct values. After the smoothing effort, we can come up with four bins, which improve the modeling tool performance as well as facilitate precise interpretation in the upcoming analysis tasks later. The bins are: 1-96, 96-191, 192-286, 286-387 as summarized in Table 3.17:

Table 3.17: Result of Discretization for OACD4

Labels	Frequencies
1--96	**2422**
96--192	**14914**
191--286	**1068**
286--387	**336**
Total	**18740**

A total of 14 attributes were selected for the research based on their relevance and preprocessing activities of the problem. There were six attributes which were excluded in the preliminary data observation like 'Item Number', Card Number, and 'Registration Date', 'Eligible Date', 'Eligible and Ready Date', 'On ART Date', since they are no more important for the mining purpose. 'Hospital' attribute was newly constructed attribute. 'Year' and 'Month-Year' were derived from registration date of each patient. The four dates above 'Eligible Date', 'Eligible and Ready Date', 'On ART Date' integrated into column since they don't hold values for all patients apart from patients who are on that particular ART stage. For example 'OA Date' has values for those patients who were on ART. Then when compared with the registration date, they are almost similar to it. Therefore; the decision was made to remove those columns from the dataset used for

model building. On top of this, the interest is on the date; on which they are registered for the care. Then the conceptual hierarchy of the registration date was segmented, to get month-year, year. This means year and month-year are derived attributes from the registration date.

Hence the final attributes used for model building are; 'Sex', 'Age', 'Functional Status' ,'Termination', 'Religion', 'Educational Level', 'ART Stage', 'Occupation', 'Marital Status', OAWHO ('WHO stage') and OACD4 ('CD4 counts') , 'Hospital', 'Year', and 'Month-Year'.

3.4 Model Implementation and Experimentation

In this section the researcher is interested to explain the model/problem selected for this research. As it is mentioned in the methodology section under 1.4 in chapter one, the problem that this research is going to address is a classification and association rule problems. Hence, it is important to explain the classification and association rule problems implementations for model building and experiments to be carried out in the data mining process, which also involve data mining tool selection and algorithms used for modeling. The classification model to be built is decision tree. The researcher is interested to experiment both binary decision tree and generalized decision tree with and without pruning. That means, there are four scenarios to be experimented for classification purpose.

For the association rule modeling purpose, the researcher implemented the frequent itemset method for finding out interesting association rule in finding what attributes and their values are frequently associated to continuity/termination behavior of patients under investigation. In finding the best frequent itemset for the association rule, the models experimented by the researcher with different parameters (minimum support and confidence thresholds) using Apriori algorithm available in weka.

Therefore it is important to discuss in brief way about decision tree classification and how it prune irrelevant tree branches to increase the classification performance of the model as well as association rule mining and how it finds out the frequent itemset in this section. Information produced by data mining techniques can be represented in many different ways of which classification and association rule are commonly motioned for predictive and descriptive data mining modeling respectively.

3.4.1 Decision Tree

Decision tree structures are a common way to organize classification schemes. In classifying tasks, decision trees visualize what steps are taken to arrive at a classification. Every decision tree begins with what is termed a root node, considered to be the "parent" of every other node. Each node in the tree evaluates an attribute in the data and determines which path it should follow. Typically, the decision test is based on comparing a value against some constant. Classification using a decision tree is performed by routing from the root node until arriving at a leaf node (Bramer, M., 2007). Bramer, M., (2007) extends his explanation of data type that can be handled by decision trees classification as:

> *Decision tree can represent diverse types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. Decision tree induction algorithms operate recursively. First, an attribute must be selected as the root node. In order to create the most efficient (i.e., smallest) tree, the root node must effectively split the data. Each split attempts to pare/trim down a set of instances (the actual data) until they all have the same classification.*

The best split is the one that provides the most information gain. Information in this context comes from the concept of entropy from information theory, as developed by Claude Shannon. Although "information" has many contexts, it has a very specific mathematical meaning relating to certainty in decision making. Ideally, each split in the decision tree should bring us closer to a classification. One way to conceptualize this is to

see each step along the tree as removing randomness or entropy (Han, J. and Kamber, M., 2006).

The process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class). The 'entropy method' of attribute selection is to choose to split on the attribute that gives the greatest reduction in (average) entropy, i.e. the one that maximizes the value of information gain. At any stage of this process, splitting on any attribute has the property that the average entropy of the resulting subsets will be less than (or occasionally equal to) that of the previous training set (Bramer, M., 2007).

The entropy of the training set is denoted by E. It is measured in 'bits' of information and is defined by the following formula as presented in Bramer, M. (2007). The following explanation of using entropy based information gain for tree pruning in decision tree is made the basis on Kamber and Han, (2006). Let node N represents or holds the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by:

$$Info(D) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Where p_i is the probability that an arbitrary tuple in D; belongs to class C_i and is estimated by $|C_i, D| / |D|$. A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. At this point, the information we have is based solely on the proportions of tuples of each class. Info(D) is also known as the entropy of D.

Suppose we were to partition the tuples in database D on some attribute A having V distinct values, $\{a_1, a_2, \ldots, a_v\}$ as observed from the training data. If A is discrete-valued, these values correspond directly to the V outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, $\{D_1, D_2 \ldots D_V\}$; where D_j contains those tuples in D that have outcome a_j of A. These partitions would correspond to the branches grown from node N. Ideally; we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). The amount of information we would still need (after the partitioning) in order to arrive at an exact classification is measured by:

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} * Info(D_j)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the j^{th} partition. $Info_A(D)$ is the expected information required to classify a tuple from *D* based on the partitioning by *A*. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on *A*). That is,

$$Gain_{(A)} = Info_{(D)} - Info_A(D)$$

Gain $_{(A)}$ tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain (A)), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to

finish classifying the tuples is minimal (i.e., minimum $Info_A(D)$).

So the researcher need to implement the entropy based attribute subset selection for tree pruning in this particular research. Weka data mining tool selected for decision tree model building has J48 implementation which uses information gain method for tree pruning. Kamber and Han, (2006) explained the advantage of information gain approach for tree pruning in decision tree over gini index for example as “ Information gain, do not force the tree branching to be binary as in Gini index, therein allowing multi-way splits (i.e., two or more branches to be grown from a node”.

J48 is a version of an earlier algorithm developed by J. Ross Quinlan, the very popular C4.5. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. It is important to understand the variety of options available when using this algorithm, as they can make a significant difference in the quality of results. In many cases, the default settings will prove adequate, but in others, each choice may require some consideration. The J48 algorithm gives several options related to tree pruning. Many algorithms attempt to "prune", or simplify, their results. Pruning produces fewer, more easily interpretable results. More importantly, pruning can be used as a tool to correct for potential over fitting. The basic algorithm described above recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible (Han J, Kamber M, 2006).

This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular idiosyncrasies of that data. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data. This is because pruning employs various means to relax the specificity of the decision tree, hopefully improving its performance on test data. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy (Witten IH, Frank E, 3005). J48 in weka 3.6.0 employs two pruning methods. The first is known as subtree replacement. This means that nodes in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The

second type of pruning used in J48 is termed sub tree raising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtree raising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtree raising can be somewhat computationally complex (Weka manual, 2008).

Error rates are used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential overfitting. This approach is known as reduced-error pruning**.** Though the method is straight-forward, it also reduces the overall amount of data available for training the model. For particularly small datasets, it may be advisable to avoid using reduced error pruning.

Other error rate methods statistically analyze the training data and estimate the amount error inherent in it. There are several other options that determine the specificity of the model. The minimum number of instances per leaf is one powerful option. This allows you to dictate the lowest number of instances that can constitute a leaf. The higher the number, the more general the tree is. Lowering the number will produce more specific trees, as the leaves become more granular. The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using an inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees. There is also an option available for using Laplace smoothing for predicted probabilities. Laplace smoothing is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities. The most basic parameter is the tree pruning option. If you decide to employ tree pruning, we may need to consider the options for pruning. It is important to know that depending on how the training and test data have been defined that the performance of an unpruned tree may superficially appear better than a pruned one (Weka Manual, 2008).

It is also important to experiment with models by intelligently adjusting these parameters. Often, only repeated experiments and familiarity with the data will tease out the best set of options as shown bellow in Table 3.18 as taken from Weka Manual (2008).

Table 3.18: Description of J48 classifier Parameter Options in Weka

Option		Description
binarySplits		Whether to use binary splits on nominal attributes when building the trees.
confidenceFactor		The confidence factor used for pruning (smaller values incur more pruning).
debug		If set to true, classifier may output additional info to the console.
minNumObj		The minimum number of instances per leaf.
numFolds		Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.
reducedErrorPruning		Whether reduced-error pruning is used instead of C.4.5 pruning.
saveInstanceData		Whether to save the training data for visualization.
seed		The seed used for randomizing the data when reduced-error pruning is used.
subtreeRaising		Whether to consider the subtree raising operation when pruning.
unpruned		Whether pruning is performed.
useLaplace		Whether counts at leaves are smoothed based on Laplace.

Since Weka data mining tool implements J48 classifier with different parameter, the researcher implemented J48 algorithm for applying decision classification model on the ART dataset preprocessed as in the previous section of the chapter.

3.4.2 Implementation of Association Rule

Association Rule Mining algorithms need to be able to generate rules with confidence values less than one. However the number of possible association rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. In an association rule mining, computational efficiency is greater concern since each of the attributes can appear with any of its possible values and any where in the association rule. Therefore; there might be a very large number of possible rules. Generating all of these is very likely to involve a prohibitive amount of computation, especially if there are a large number of instances in the dataset. To identify the best quality of association rules, we need to see a kind rule interestingness measures (Bramer, M., 2007).

Several interestingness measures have been proposed in the technical literatures. In this research the researcher followed the one proposed in Bramer M. (2007) to describe how interestingness measures of association rule is measured. In association rule mining, the rules appear in the form of:

If Antecedent Then Consequent

We start by defining four numerical values which can be determined for any rule simply by counting:

Let; $N_{Antecedent}$ be the number of instances matching Antecedent. $N_{consequent}$ be the number of instances matching Consequent N_{BOTH} be the number of instances matching both Antecedent and Consequent N_{TOTAL} be the total number of instances

These numerical values can be represented using Venn diagram to visualize the concept as in figure below

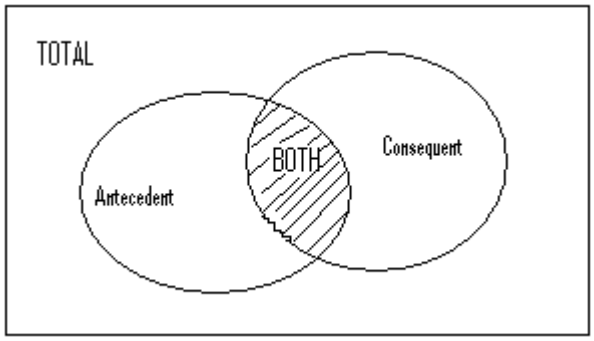


Fig. 3.2: Instances Matching Antecedent, Consequent and both

Measures of rule interestingness (support and confidence mentioned under Chapter 2), can be computed from these four numerical values. The proportion of both Consequent and antecedent occurring together to the occurrence of the antecedent is called confidence and it is computed as:

Confidence (Rule Accuracy, Reliability) = N_{BOTH} / Antecedent

The proportion of the training set correctly predicted by the rule is called support and computed as:

Support (Prevalence) = N_{BOTH} / N_{TOTAL}

In general; association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data (transactions).The rules are written as X=>Y, where X is called the antecedent or Left-Hand Side (LHS), and Y is called the consequent or right hand side (RHS) (Bramer M, 2007).

Association algorithms find these rules by doing the equivalent of sorting the data while counting occurrences so that they can calculate confidence and support. The efficiency with which they can do this is one of the differentiators among algorithms. We should be able to evaluate rules using different techniques especially because of the combinatorial explosion that results in enormous number of rules and for the fact that all the rules are not interesting. As written in Han and Kamber (2006), association rules are considered

interesting if they satisfy both minimum support and confidence thresholds. Rules that satisfy both a minimum support threshold and a minimum confidence threshold are strong rules. For the implementation of the association rule in this research, Appriori algorithm, which is available in weka data mining tool to find frequent itemset, is used.

CHAPTER FOUR

MODEL BUILDING

4.1 Attribute Ordering

Since attribute selection is important in decision tree models, the researcher tried to rank the attribute based on information gain, which was thoroughly discussed in Chapter 3, section 3.4.1. Ranking the attributes to the mining task of the decision tree was implemented by Weka attribute ranking filter using information gain.

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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 Termination):

        Information Gain Ranking Filter

        Ranked attributes:

0.0382893   4     ARTStage
0.0275886   7     FunctionalStatus
0.0252408   5     Month-Year
0.0170153   6     Year
0.0150237  13     OACD4
0.0077979  12     OAWHO
0.0020178   3     Age
0.0014028   2     Sex
0.0007097   8     MaritalStatus
0.0004846  11     Occupation
0.000219   10     Religion
0.0001606   9     EducationalLevel
0.0000716   1     Hospital

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

As you can see from the result of attribute selection using entropy based information gain method of weka, the top six determining attributes of the dataset for predicting

continuity/termination behavior of ART care service by the patients are ART Stage, Functional Status, Month-Year, Year, OACD4 (CD4 count and OAWHO (WHO Stage) of HIV/AIDS with information gain of 0.038, 0.028, 0.025, 0.017, 0.015 and 0.0077 respectively. The numbers followed the selected attributes: - are the attributes' indices in the dataset and the number 13 shows total number of attributes ordered in determining the class label "Termination", which is the 14th attribute. Knowing attributes' relevance to the data mining task as depicted above helps for later experimentation by excluding the least relevant attributes like hospitals turn by turn.

4.2 Building Classification Models

Four scenarios were intended to be built as per stated in the discussion of model selection and implementation in Chapter three section 3.4.1. These are binary decision tree with and without pruning and generalized decision tree with and without pruning with all attributes. Trying those four scenarios with top 10 attributes (reduced attributes) doubles the number of the scenarios and the researcher ends up in experimenting eight scenarios.

These classification models are discussed in this section of the chapter to compare their efficiency of the classifier models so that applicability of decision tree classification models will be discovered to the domain problem as stated in the objective section of chapter one. To see them in a more detail, a decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes. Decision trees are commonly used in decision analysis, to help identify a strategy most likely to reach a certain goal. They are one of the most widely used and practical forms of machine learning and data mining. They have been widely researched and applied to a large variety of data mining problems. Decision tree models are built by a process that is known as recursive partitioning (Bramer, M., 2007).

In building the tree models, first, the original data is broken up into 2 or more non-overlapping sub-samples. The original sample is also known as the root node and each of the sub-samples is referred to as a node. The partitioning is done based on one of the independent variables known as the splitting attribute. Branches are drawn for different values of this splitting attribute. Each instance in the root node is sent down one of the

branches (depending on its value for the splitting attribute) into one of the nodes. The choice of splitting attribute is done by picking the attribute that will partition the original sample into sub-samples that are as homogenous as possible in relation to the class variable. It is a similar idea when choosing on what values of the splitting attribute to perform the partitioning.

As a second step, the process is repeated for each of the nodes created in the first step. Each node is partitioned by considering only the cases in that node and new nodes are created from instances in that node alone. This process is repeated for each node until some stopping-rule is violated. When this happens, the node is not further partitioned and this node is referred to as leaf node. The whole process is terminated when there are only leaf nodes left.

The ultimate goal of building a tree model is to end up with the smallest tree that has the purest leaf nodes. The purer a leaf node is, the more precise its classification is. A leaf node where all the instances in it are correctly classified is clearly superior to one where just of over half are correctly classified. Different algorithms will tend to behave better or worse on some data as opposed to other data.

So, it is important that we select algorithm for the implementation of decision tree modeling and how it prunes for irrelevant branches of the tree. For this purpose you can see the details of model selections and implementation to section 3.2.1 of Chapter 3. The goal of splitting up a sample is to get sub-samples that are more pure than the original sample. Purity in this case is referring to how homogenous the sub-samples are in relation to the class variable. The ideal situation is when each sub-sample consists of instances that have the same value for the class attribute i.e. completely pure nodes (Kamber and Han, 2006). Before building decision tree classification model; the researcher believed that it is good to decide on the portion of the dataset, which is used for training and testing purpose. To do so learning curve is found important to get initial perception about the dataset at hand.

4.2.1 Selection of Validation Method for Decision Tree Models

In classification model building, validation (measuring classifier accuracy on unseen data) is very important issue. There are various validation methods for decision tree classification problem for the model accuracy. The types of validation methods include dividing the data into a training set and a test set (full training set), k-fold cross-validation (10-fold cross validation in weka) and N-fold (or leave-one-out) cross-validation. Therefore it is important to check the appropriateness of dataset for selecting certain validation method (Bramer, M. 2007). To opt for one of these validation methods, it is important to see its appropriateness for the available dataset.

To see the appropriateness of the dataset preprocessed and ready for model building above in section 3.2, the researcher tried to observe its learning curve. From the graph on the learning curve, the point at which the learning curve converges might show the minimum of sample dataset to be used for training purpose so that the rest will be planned for testing purpose. If the learning curve, which can show the maturity/saturation point goes sharply up as the sample size increases then 10-fold cross validation will be used to minimize the effect of the shortage of the dataset used. For this research purpose the experiment started from 10% gone through up to 90% of the dataset as shown on the learning curve. For illustration, it was presented on figure 4.1. To make sure that the learning curve revealed correct point of convergence to decide splitting point of the training and testing dataset, the experimentation of 10- fold cross validation is made and it showed the same performance with that of the full training set at 50% split for training dataset. The learning curve is drawn only based on the generalized decision tree with pruning model with all attributes. The percentage of the samples and the performance of the classifier at different sample dataset are presented in Table 4.1.

Table 4.1: Samples of Training Dataset and Corresponding Performance of Classifier

Sample	Performance
10%	74
20%	76
30%	78
35%	78
45%	78
50%	78
55%	78
65%	78
75%	78
85%	78
90%	78

When we look at the learning curve, the convergence has happened at 35% of the dataset. The performance of the classifier goes uniformly after 35%. The researcher decided to use 50% (9370) of the dataset for training and the remaining 50% for testing dataset to be fair for other scenarios and sufficient testing will be available even though the learning curve converges at 35%. The learning curve looks like as in the figure 4.2:

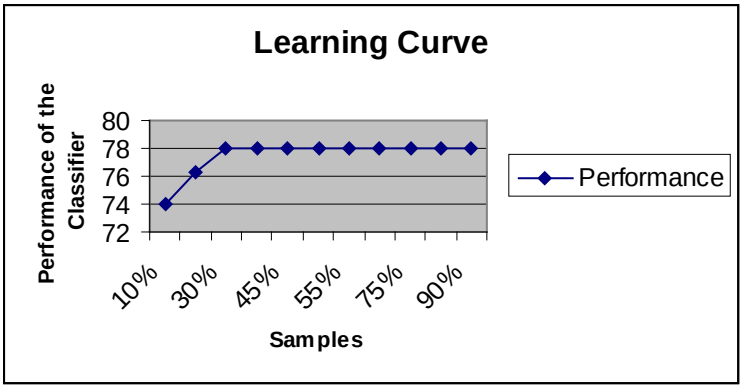


Fig. 4.1: Learning Curve for Training Dataset

4.2.1 Building Binary Decision Tree

As it has been mentioned under section 3.4.1, one of the scenarios to be implemented for decision tree classification in this research is binary decision tree. Binary trees split internal node branches to exactly two sub trees) or we will have two branches at each node (Bramer M., 2007). Hence binary decision tree with and without pruning using J48 classifier in weka is built.

A binary tree is a tree with one more restriction: no node may have more than 2 children as compared to that the generalized tree one when a node can carry a minimum of two and maximum of many children (Bramer, M, 2007). The binary decision tree scenarios to be experimented by the researcher are: Binary decision tree with all attributes without pruning, binary decision tree with all attributes with pruning, binary decision tree with some selected attributes without pruning, binary decision tree with some selected attributes with pruning.

4.2.2 Modeling Generalized Decision Tree

Generalized decision tree model is built by setting the 'binarySplit' to 'False' so that a single node can be splited into more than two sub trees. But those tree branches may be of less importance to reveal important and concise knowledge as discussed above in decision tree building section.

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of overfitting the data. Such methods typically use statistical measures to remove the least reliable branches. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data (i.e., of previously unseen tuples) than unpruned trees (Kamber and Han, 2006).

Therefore the researcher built generalized decision tree to compare its efficiency as compared to other scenarios experimented in building the decision tree classification model. There are also four scenarios for generalized decision tree experimented in this research. These are generalized decision tree with all attributes with pruning, generalized decision tree with all attributes without pruning, generalized decision tree with some of the attributes with pruning and generalized decision tree with some of the attributes without pruning.

Weka has J48 implementation for all the decision tree scenarios discussed above. For understanding the explorer window look at Fig. 4.2.

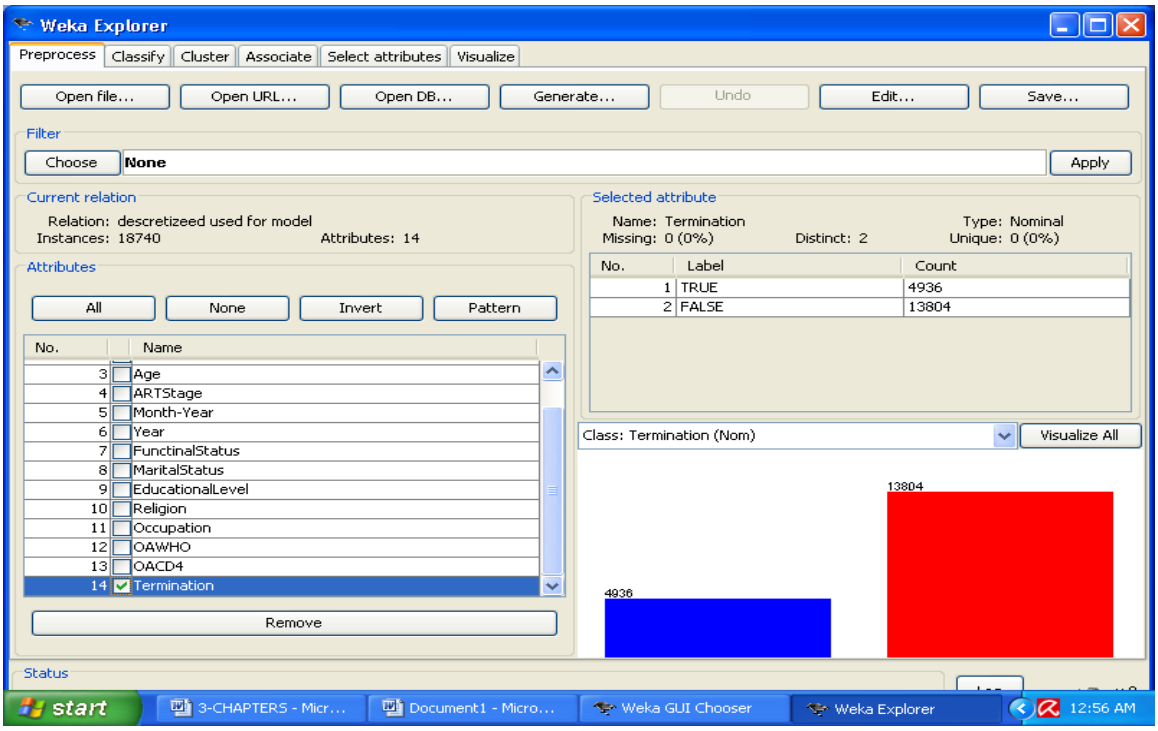


Fig. 4.2 Weka explorer window

The J48 classifier window which enables us to switch the parameters to build different decision tree scenarios is indicated in Fig. 4.2.

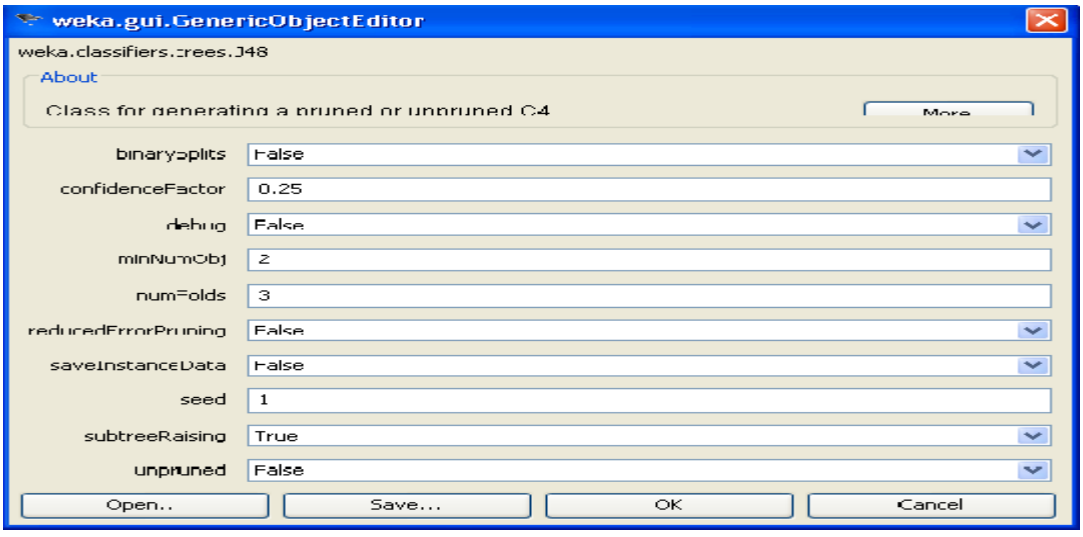


Fig. 4.3: Weka Parameters Window for J48 Classifier

For classification purpose with implementation J48 classifier, parameters are changed from the weka Generic ObjectEditor window as indicated on figure 4.3 for building different decision trees. Meanings of the parameters on this window were already given in chapter 3 under section 3.4.1 in table 3.18. It is from this window that whether to built binary or generalized decision tree is decided. The parameter used for this purpose is 'binarySplit'; which has 'True' and 'False' values. Setting the parameter to 'True' enables us to build binary decision tree. When we set the parameter to 'False', then the resulting result is generalized decision tree. Another important parameters on this window is relevant to this research is 'unpruned', again has 'True' and 'False'. Setting 'unpruned' to 'True' enable us tell the classifier we don't want prune the tree setting it to 'False' is to mean we want to prune.

4.3: Building Association Rule Model

Association rules were the second data mining technique to be used for this research. There are different algorithms available for implementing association rule mining. Apriori algorithm is the most commonly used association rule mining algorithm and also selected for this research. The researcher tries to mine important (interesting) rules at different minimum support and confidence thresholds using weka as a tool. One of the most commonly used association rule modeling technique is apriori algorithms.

Apriori is an algorithm proposed in Agrawal, R. and Srikant R. (1994) for mining frequent itemsets for boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see in the following discussion. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L_1. Next, L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.

Not all rules are interesting. Those rules which satisfy the minimum support and confidence thresholds are interesting. Therefore; the researcher experimented different association rule mining using apriori algorithm at different minimum support and confidence thresholds on weka.

General information regarding association rule mining and how it finds out the best rules based on the interestingness measures (minimum support and minimum confidence thresholds) was presented in Chapter three sections 3.4.2. The experimentations were carried out by the researcher at different minimum support and confidence thresholds. The minimum support starts at 0.3 was experimented until 0.1 with varying the minimum confidence thresholds from 50% to 100%. The actual experiment and analysis of association rules will be made in chapter six. Weka has the association rule implementation of Apriori algorithm.

Weka window for tuning parameters for association rule is presented in Fig 4.4. This window has got different for changing parameters related to minimum support and confidence thresholds.

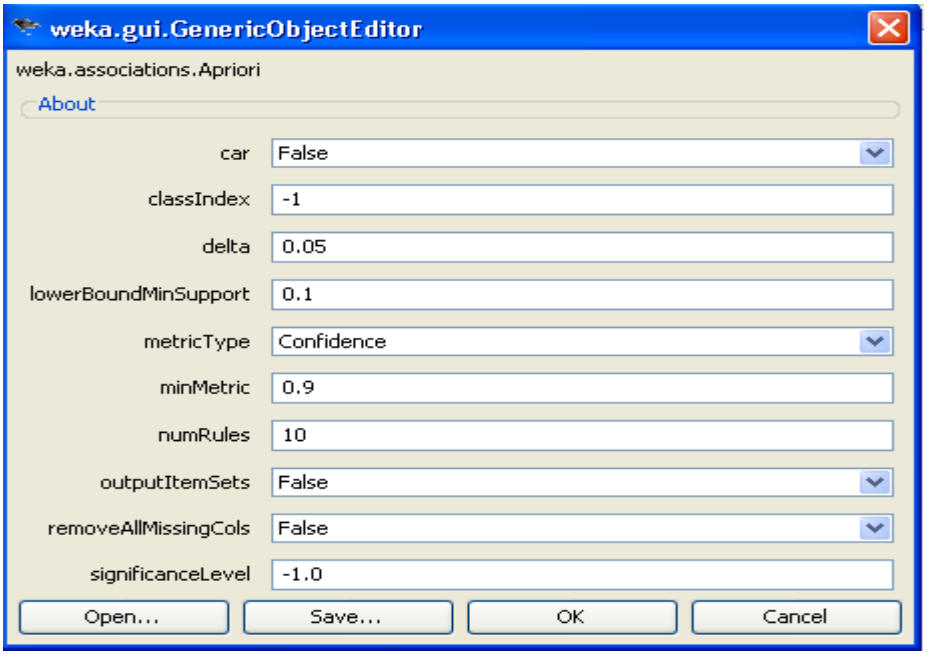


Fig. 4.4: Weka Appriori Window for Association Rules Parameters Setting

Let us see meanings of some of the parameters on this window in Table 4.1

Table 4.2: Meanings of the Parameters for Association Rule

NumRules	Required number of rules output
MinMetric	Confidence
Delta	delta at which the minimum support is decreased at each iteration
UpperbondMinSupport	Upper bound for minimum support
-LowerBondMinSupport	Lower bound for minimum support

CHAPTER FIVE

EXPERIMENT AND ANALYSIS OF CLASSIFICATION MODEL

Analysis of the decision tree models are made in terms detailed accuracy of the classifier on the training dataset as tested on the test data based on a confusion matrix of each model result. The confusion matrix is a useful tool for analyzing how well our classifier can recognize tuples of different classes (True and False classes in the case of this research). Confusion matrix shows four important numerical quantities.

As it has already been discussed in Chapter 4, there are eight scenarios to be experimented for decision tree classification. These scenarios are going to be experimented and analyzed to compare them to each other in terms of different performance matrices values, accuracies, number of leaves, and size of tree generated, ROC curves and execution time. The models are also compared with regard to the patterns/ knowledge discovered. The scenarios for decision tree classification experimented in this research are as listed bellow: we are going to see the models results and the analysis of each result and compare the result of one model to the previous one and finally to find out the outperforming model based on the criteria of evaluation.

Scenario #1: Binary Decision Tree without pruning with all Attribute
Scenario #2: General Decision Tree without pruning with all Attribute
Scenario #3: Binary Decision Tree without pruning with reduced Attribute
Scenario #4: General Decision Tree without pruning with reduced Attribute
Scenario #5: Binary Decision Tree with pruning with all Attribute
Scenario #6: General Decision Tree with pruning with all Attribute
Scenario #7: Binary Decision Tree with pruning with reduced Attribute
Scenario #8: General Decision Tree with pruning with reduced Attribute

As we have already seen in Chapter section 4.2.1, the method of validation is decided to be full training set splitting at 50% (9370) of the dataset for training and allocating the rest 50% testing dataset.

Scenario #1: Binary Decision Tree without pruning with all Attribute

The result of this scenario is as presented below. The first line reports the split point for training and testing dataset, which was already addressed in Chapter 4, section 4.1.1.

Test mode: split 50.0% train, remainder test

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

J48 unpruned tree

Number of Leaves: 2811

Size of the tree: 5621

Time taken to build model: 5.59 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	6477	69.1249 %
Incorrectly Classified Instances	2893	30.8751 %

The base for calculating Correctly Classified Instances and Incorrectly classified Instances, as it is mentioned in the methodology section is the confusion matrix. You can also see the detailed accuracy measure of this scenario as presented bellow:

	TP Rate	FP Rate	Precision	Recall	ROC Area
Class					
TRUE	0.376	0.195	0.409	0.376	0.604
FALSE	0.805	0.624	0.782	0.805	0.604
Avg.	0.691	0.511	0.683	0.687	0.604

The confusion matrix of the class which is a base for calculating accuracy measures and performance is presented below.

a	b	<-- classified as
932	1548	a = TRUE
1345	5545	b = FALSE

The number of true positives in this confusion matrix is 932 records. Those records which were predicted as 'True' class by the classifier and also happened true by when tested on the test data are (True Positives). The number of the records which were classified to the "False' class by the classifier and they are actually False as tested on the test data (True Negative Rate) is 5545. The sum of TPR (932) and TNR (5545) gives us correctly classified.

The total number of the records which were correctly classified to true and false classes of termination status of the patient was 6477.

The summary Table of the above model results can presented as follows:

Table 5.1: Summary of Scenario #1

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	*CCI	ROC Area	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
1	Binary	No	All	2811	5621	6	69%	0.60	0.69	0.51	0.68	0.69	0.38	0.81

As we can see from Table 5.1, Binary Decision Tree without pruning with all attributes scenarios has generated complex tree whose tree structure can not be visualized. It has many leaf nodes as well as it is very lengthy. Otherwise its ability in correctly classifying records into both 'True' and 'False' Classes is moderate (69%). Its ROC area is also above 0.5, which is the minimum possible acceptable value for ROC curve. If draw the ROC Curve 0.60 above the diagonal. ROC area is plotted from True positive Rate (TPR) on the y axis against the False Positive Rate (FPR) on the x-axis. So it means that there is a 50-50 chance for TPR and FPR, which unacceptable.

Based on these criteria it is possible to evaluate the performance of this model from the value of ROC area is above 0.5, which is 0.60 and we can say the model is moderately accurate. As we can see from the ROC area of True class, the curve is above the diagonal line. But as it goes up , it bends to the diagonal point.

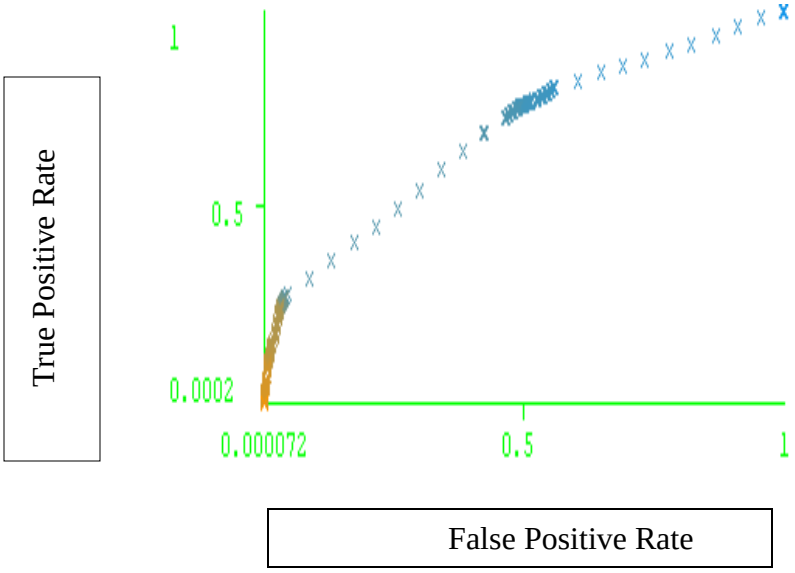


Fig. 5.1: ROC Area curve for Scenario#1

Formula for calculating detailed accuracy measures of the classifiers was discussed in the methodology section in chapter on section 1.4.1.5.

Specificity of this scenario is 80% (True positive rate for the false class) while its sensitivity is 38%. Specificity of the classifier is the ability of the classifier to identify the true negatives (Actual False in this case). Sensitivity is the ability of the classifier to identify true positives, which is equal to the recall of the classifier. Precision of this classifier on average is 0.68. The precision of this model for the false class is higher than the precision for the true class.

Scenario #2: General Decision Tree without Pruning with All Attribute

The result of this Model is as follows;

Test mode: split 50.0% train, remainder test

Number of Leaves: 8594

Size of the tree: 10082

Time taken to build model: 1.92 seconds

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

Correctly Classified Instances 6561 70.0213 %
Incorrectly Classified Instances 2809 29.9787 %
Total Number of Instances 9370

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	ROC Area
TRUE	0.398	0.191	0.429	0.412	0.615
FALSE	0.809	0.602	0.789	0.809	0.615
Avg.	0.7	0.493	0.693	0.7	0.615

=== Confusion Matrix ===

a	b	<-- classified as
986	1494	a = TRUE
1315	5575	b = FALSE

Analysis of this scenario is made as per sated in the methodology section of Chapter one. The number of correctly classified instances 6561(70%). This means that the number of records which are correctly classified to both the 'True' and 'False' classes of the termination status of the patients while they are actually in those classes when tested on the test data.

Compared the performance, Accuracy measures and other matters with first Scenario, it is better. Let us see in terms of number of leaves and size of the tree for scenario #1 and #2, time taken, correctly classified instances (CCI in the table) and ROC area.

Table 5.2: summary of performance of Scenario #2

Scenarios#	No of leaves	Size of Tree	Time(Sec)	CCI	ROC Area	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
2	8594	10082	2	70%	0.62	0.70	0.49	0.69	0.70	0.40	0.81

Both number of leaves and size of the tree are greater for scenario two which is general decision tree without pruning with all attributes. This shows that the second scenario is more complex. But scenario #2 is better than scenario #1 in terms of execution time, CCI, and ROC area. When we compare both scenarios in terms of TPR, Recall, Precision,

Sensitivity and specificity, FPR is less for scenario #2, better than scenario #1. TPR, Precision, Recall and Sensitivity are higher for the general decision tree without pruning with all tribute scenario compared to scenario #1.

Scenario #3: Binary Decision Tree without pruning with reduced Attribute

Some of the attributes excluded in those scenarios in which reduced attributes appears in this research are those attributes which came on the last four ranks in attribute ordering in Chapter four, section 4.1. As you may remember these attributes are 'Occupation', 'Religion', 'Educational level' and 'Hospital'. So we are going to have four scenarios without these attributes together with Binary decision tree with and without pruning and generalized decision tree with and without pruning. The result of Scenario three looks like as follows:

Test mode: split 50.0% train, remainder test

Number of Leaves: 1836

Size of the tree: 3671

Time taken to build model: 3.72 seconds

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

Correctly Classified Instances	6683	71.3234 %
Incorrectly Classified Instances	2687	28.6766 %
Total Number of Instances	9370	

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	ROC Area
TRUE	0.327	0.148	0.443	0.327	0.613
FALSE	0.852	0.673	0.779	0.852	0.613
Avg.	0.713	0.534	0.69	0.713	0.613

=== Confusion Matrix ===

a	b	<-- classified as
812	1668	a = TRUE
1019	5871	b = FALSE

In this scenario the TPR for the True class is a bit improved than the first two scenarios and less than that of the second one. Correctly classified instances (CCI) are also improved than the previous scenarios. Summary of performances of the third scenario is presented in Table 5.4.

Table 5.4: Summary of scenario #3

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	CCI	ROC Area	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
3	Binary	No	Reduced	1836	3671	4	71%	0.61	0.71	0.53	0.69	0.71	0.33	0.85

When we compare scenario #3 with previous two models the number of leaves decreased significantly, the size of the tree also become smaller and correctly classified instances are getting increased. The time taken in building the model three is more than that of the second scenario and less than the first scenario as indicated in Table 5.4. We can see from this Table that ROC Area is of the third classifier model is less than that of the second model and greater than the first model. This means the second scenario is more performing identifying True Positives as it is the case for the third model compared to the first model. FPR is higher than the previous scenarios.

Scenario #4: General Decision Tree without pruning with reduced Attribute

The fourth scenario build with parameter of general decision tree without pruning with some attributes. What it means by reduced attributes has already explained in the third scenario. The important summary of this result of the model is as follows:

Number of Leaves: 1836

Size of the tree: 3671

Time taken to build model: 3.72 seconds

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

Correctly Classified Instances 6683 71.3234 %
Incorrectly Classified Instances 2687 28.6766 %
Total Number of Instances 9370

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	ROC Area
TRUE	0.327	0.148	0.443	0.327	0.613
FALSE	0.852	0.673	0.779	0.852	0.613
Avg.	0.713	0.534	0.69	0.713	0.613

=== Confusion Matrix ===

a	b	<-- classified as
812	1668	a = TRUE
1019	5871	b = FALSE

Summary of this model is presented in Table 5.6.

Table 5.6 Summary Performance of Scenario #4

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	CCI
4	General	No	reduced	1836	3671	4	71%

As you can see from Table 5.6, the number of leaves, size of the tree, time taken by the third and fourth model is identical. Therefore comparison made among the first three scenarios can work for the first two models. The rest of the accuracy measures of the fourth scenario are also similar to the third model. Therefore binary and general decision tree without pruning with reduced attributed selected based on attribute ranking can equally perform on the ART dataset but better than first two scenarios except greater execution time compared the second model.

Scenario #5: Binary Decision Tree with pruning with all Attribute

The fifth classification model is built with parameter of Binary decision Tree with pruning with all attributes. Its result is presented bellow. The researcher tried to compare the result of this model to the previous models in terms of all matters pertaining to the

performance, detailed accuracy, number of leaves, size of tree etc… after the model result

Number of Leaves: 297

Size of the tree: 593

Time taken to build model: 25.17 seconds

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

Correctly Classified Instances	7077	75.5283 %
Incorrectly Classified Instances	2293	24.4717 %
Total Number of Instances	9370	

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	ROC Area
TRUE	0.288	0.077	0.575	0.288	0.653
FALSE	0.923	0.712	0.783	0.923	0.653
Avg.	0.755	0.544	0.728	0.755	0.653

=== Confusion Matrix ===

a	b	<-- classified as
715	1765	a = TRUE
528	6362	b = FALSE

Performance summary of Scenario #5 is presented in Table 5.7

Table 5.7: Performance Summary of Scenario #5

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	CCI	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
5	Binary	yes	All	297	593	25	76%	0.76	0.54	0.73	0.76	0.29	0.92

As you can see from Table 5.7, everything gets butter in the fifth model except taking more execution time. The tree complexity is reduced relatively compared to the previous

ones. Correctly classified instances are also significantly increased. When we compare by the detailed accuracy measures, the same thing happens. The figures in the Table for the fifth model are greater on any of the criteria to measure performance the model than the previous models.

Scenario #6: Generalized Decision Tree with pruning with all Attribute

The researcher built generalized decision tree to compare its efficiency to other scenarios experimented so far. Hence the researcher preferred to present the full model of this scenario in the body of the research. Models result of other decision tree scenarios were annexed at the end due to their complexity.

```
Test mode:    split 50.0% train, remainder test

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

J48 pruned tree
------------------

ARTStage = OA
|   FunctionalStatus = A
|   |   Year <= 1999
|   |   |   Age = '(31.5-44.75]': TRUE (355.0/113.0)
|   |   |   Age = '(44.75-inf)': TRUE (139.0/39.0)
|   |   |   Age = '(18.25-31.5]': TRUE (370.0/118.0)
|   |   |   Age = '(-inf-18.25]': FALSE (37.0/7.0)
|   |   Year > 1999
|   |   |   Hospital = Adama: FALSE (484.0/97.0)
|   |   |   Hospital = Asella
|   |   |   |   Year <= 2000: TRUE (184.0/65.0)
|   |   |   |   Year > 2000: FALSE (205.0/85.0)
|   FunctionalStatus = W: FALSE (8026.0/2249.0)
|   FunctionalStatus = B
|   |   Year <= 1999: TRUE (537.0/91.0)
|   |   Year > 1999
|   |   |   Hospital = Adama: FALSE (167.0/45.0)
|   |   |   Hospital = Asella: TRUE (84.0/19.0)
|   FunctionalStatus = P: FALSE (597.0/122.0)
|   FunctionalStatus = D: FALSE (67.0/23.0)
|   FunctionalStatus = R: FALSE (10.0/4.0)
ARTStage = IN: FALSE (7037.0/973.0)
ARTStage = ER: FALSE (25.0/3.0)
ARTStage = EL: FALSE (416.0/104.0)
```

Fig. 5.1: Model structure of General Decision Tree with Pruning and with all attributes

```
Number of Leaves:       17

Size of the tree:       25


Time taken to build model: 2.52 seconds

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

Correctly Classified Instances        7264               77.524 %
Incorrectly Classified Instances      2106               22.476 %
Total Number of Instances             9370

=== Detailed Accuracy By Class ===

        Class   TP Rate    FP Rate   Precision      Recall     ROC Area
                 0.242     0.033      0.726         0.242         0.682
TRUE
FALSE            0.967      0.758     0.78          0.967         0.682
Avg.             0.775      0.566     0.766         0.775         0.682
```

```
=== Confusion Matrix ===

    a    b   <-- classified as
  600 1880 |    a = TRUE
  226 6664 |    b = FALSE
```

Summary of the performance measures of scenario #6 is presented in Table 5.8

Table 5.8: Performance summary of Scenario#6

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	CCI	ROC Area	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
6	General	yes	All	17	25	3	78%	0.68	0.78	0.57	0.77	0.78	0.24	0.97

You can see from Table 5.8, how the sixth model (scenario) well performing on the dataset in all matters compared to the previous ones. Number of leaves and size of the tree of the tree is coming down very sharply, which shows the preciseness of this scenario compared to others. All of the detailed accuracy measures indicate the higher performance of generalized decision tree with pruning with all attributes.

Scenario #7: Binary Decision Tree with pruning with Reduced Attribute

Binary decision tree with pruning with some attributes is experimented if in case some improvement might happen by excluding the least relevant attributes. The result of model is looks as follows:

```
Number of Leaves:       144

Size of the tree:       287

Time taken to build model: 10.52 seconds

                    === Evaluation on test split ===

Correctly Classified Instances        7154               76.3501 %
Incorrectly Classified Instances      2216               23.6499 %
Total Number of Instances             9370
=== Detailed Accuracy by Class ===
```

```
Class
               TP Rate    FP Rate   Precision    Recall      ROC Area
TRUE           0.248      0.051       0.637      0.248       0.657
FALSE          0.949      0.752       0.778      0.949       0.657
Avg.           0.764      0.566       0.741      0.764       0.657
```

The confusion matrix of the model looks like as bellow:

```
    a    b   <-- classified as
  615 1865 |    a = TRUE
  351 6539 |    b = FALSE
```

The summary of performance measures of the seventh scenario is given in Table 5.9

Table 5.8 summary performance of Scenario #7

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	CCI	ROC Area	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
7	Binary	yes	reduced	144	287	11	76%	0.66	0.76	0.57	0.74	0.76	0.25	0.95

It can be seen from the Table 5.9 that the seventh scenario is less precise than its general decision tree counterpart; but better than scenarios # 1 to #5. The time it took is the largest next to binary decision tree with pruning with pruning with all attributes, which is scenario five. Taking some of the attributes for the binary tree improved its understandability and it also reduced the execution time more than with all attributes counter part (Scenario #5). When we see this scenario in terms of detailed accuracy measures, it is less performing in every aspect than the sixth scenario but better than other scenarios. But the average FPR is similar to that of the six scenarios.

Scenario #8: General Decision Tree with pruning with some Attribute

This scenario is experimented for incase some improvement might occur due to the exclusion of least relevant attributes by attribute ordering. The complexity of the tree increased at can be observed from the number of leaves and size of the tree compared to its all attribute counterpart.

Number of Leaves: 131

Size of the tree: 148

Time taken to build model: 1.3 seconds

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

Correctly Classified Instances	7213	76.9797 %
Incorrectly Classified Instances	2157	23.0203 %
Total Number of Instances	9370	

=== Detailed Accuracy by Class ===

Class	TP Rate	FP Rate	Precision	Recall	ROC Area
TRUE	0.206	0.027	0.732	0.206	0.675
FALSE	0.973	0.794	0.773	0.973	0.675
Avg.	0.77	0.591	0.762	0.77	0.675

Confusion Matrix of model is bellow:

a	b	<-- classified as
510	1970	a = TRUE
187	6703	b = FALSE

This is last experiment of the classification model and hence is good to compare with previous experiments using different performance outcomes and accuracy measures as well as complexity of the model following the result. Table 5.14 summarizes the entire summary of the eight experiments performed for classification model and it also helps to compare the last experiment with the previous ones.

Table 5.9: Al l summary of measures of performance and accuracy of the models used for comparison

Scenarios#	Type	Pruning	Attributes	No of leaves	Size of Tree	Time(Sec)	CCI	ROC Area	AVG TPR	AVG FPR	Precision	Recall	Sensitivity	Specificity
1	Binary	No	All	2811	5621	6	69%	0.60	0.69	0.51	0.68	0.69	0.38	0.81
2	General	No	All	8594	10082	2	70%	0.62	0.70	0.49	0.69	0.70	0.40	0.81
3	Binary	No	Reduced	1836	3671	4	71%	0.61	0.71	0.53	0.69	0.71	0.33	0.85
4	General	No	Reduced	1836	3671	4	71%	0.61	0.71	0.53	0.69	0.71	0.33	0.85
5	Binary	yes	All	297	593	25	76%	0.65	0.76	0.54	0.73	0.76	0.29	0.92
6	**General**	**yes**	**All**	**17**	**25**	**3**	**78%**	**0.68**	**0.78**	**0.57**	**0.77**	**0.78**	**0.24**	**0.97**
7	Binary	yes	Reduced	144	287	11	76%	0.66	0.76	0.57	0.74	0.76	0.25	0.95
8	General	yes	Reduced	131	148	1	77%	0.68	0.77	0.59	0.76	0.77	0.21	0.97

The performance of scenario # 8, in all aspects is next to the performance experiment six. The eighth scenario took smallest time of all the scenarios. The researcher also believes that it good to visualize through line graph different performance criteria's used to analyze and evaluate the models for easy understanding. Fig. 5.1 shows the trend in complexity of the trees in terms of number of leaves and size of the trees for every eight scenarios experiments. The graph shows that the maximum number of tree and biggest

size of the tree is experiment 2 and the least is experiment 6.

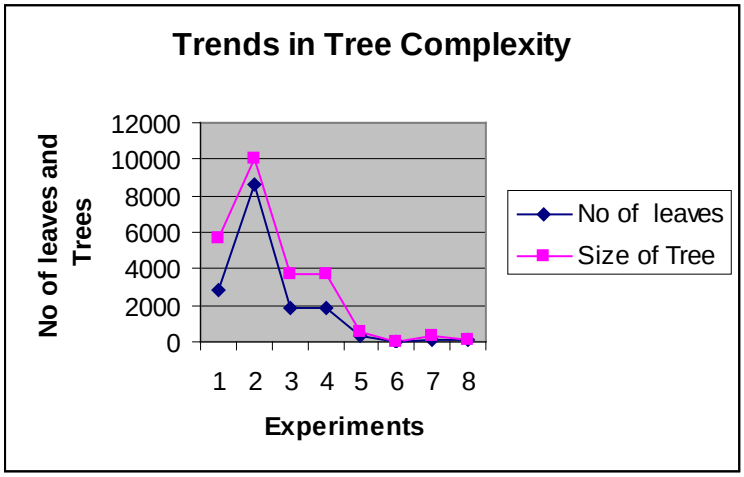


Fig. 5.2: Graph for Tree Complexity

Another important criteria used to compare the decision models above was the execution time. The execution time of all of the scenarios experimented above is presented graphically as in Fig. 5.3

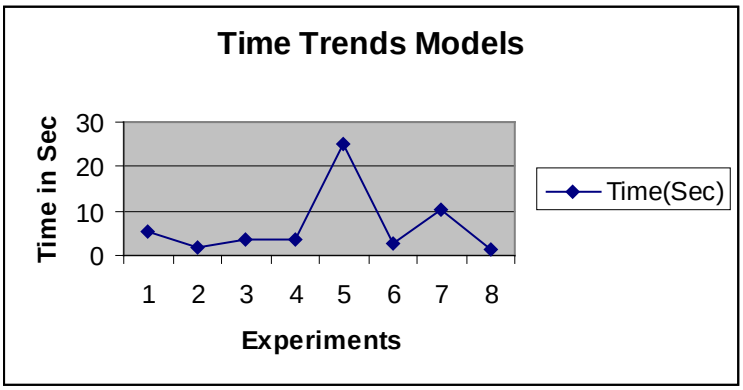


Fig. 5.3: Time Trends for all models

This graph shows that the maximum time in second is for experiment 5 the smallest time is for experiment 8. Fig. 5.3 compares the performance of the models in correctly classifying instances. It can be seen from the graph .The minimum performance observed on experiment (scenario) #1 and the maximum is for the scenario # 6 followed by scenarios.

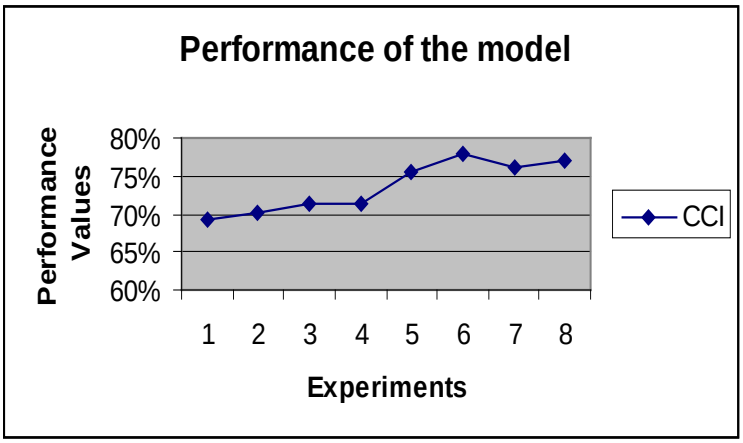


Fig. 5.4: Comparison by performance for all models

Fig. 5.5 shows the summery of the accuracy measure of all the models on a single graph. As can see from the graph, specificity is a bit high for all models while sensitivity is low across all models. Average TPR is also good for all models. The ROC area of all of the models is also above 0.5 is moderately good

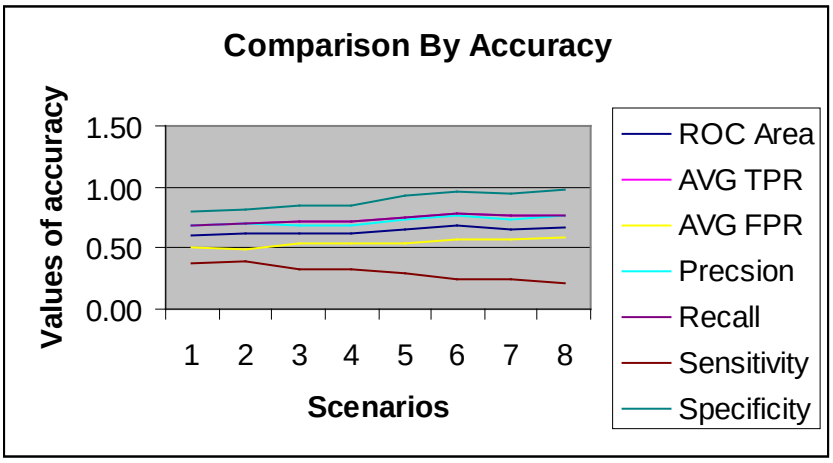


Fig. 5.5 comparison of the models by accuracy measures

Important Patterns Obtained from classification models include as follows:

It is good to see the meaning of the patterns generated by decision tree. The number of instances for each label is given at the leave as Name of the majority class follows by number of instances for the majority class/ number of the minority class in brackets. It is possible to calculate the likelihood predictability of the majority class from the numbers of instances. E.g. TRUE (355.0/113.0) is first leaf of the decision tree in the generalized

decision tree with pruning using all attributes model, in which ‘True’ value is majority class. The following are few of the patterns which were discovered between termination status and other attributes and their values. These patterns have also got acceptance by domain experts as consulted informally.

If ART is ‘OA’ (on ART) and Functional status is ‘A’ (Ambulatory) and ‘Year’ in which the patient is registered for the care is before or within 1999 and Age is between 31 to 45 years then Termination status is predicted as ‘True’ (will terminate) highly likely.

If ART Stage is ‘OA’ and Functional status is ‘A’ and year in which the patient started treatment is before or within 1999 E.C and whose Age is between 45 and 58 years then the termination status is 78% highly likely to be predicted as True (terminate the care/treatment).

If ART Stage is ‘OA’ and Functional Status is ‘W’ (Working) then the termination status is highly likely to be predicted as false (will not terminate the treatment of the ART care).

If ART Status is ‘OA’ and Functional Status is ‘B’ (bedridden) and Registration year is before or within 1999 E.C then The termination status is predicted with likely hood of 85% as ‘True’ (the patient terminates the treatment). If ART stage is ‘IN’ (in care) or didn’t start taking the Antiretroviral (ARV) drug but in care for treatment of other opportunistic infections then termination status is predicted as they will not terminate the care with the likely hood of 87%.

More generally it also indicated that if the ART stage of the patient is “OA’ (started ARV drug), and the functional status is ‘A’ (Ambulatory) and the year in which she/he is registered for the service was before or in 1999 E.C and whose age is above 18 years the then he she most likely terminates the ART Treatment.

If ART stage is OA and Functional is Ambulatory and the year in which the individual patient started the service is after 1999 E.C and Hospital is Adama then the patient will not terminate the ART care.

If ART stage is OA and Functional status is Ambulatory and and Hospital is Asella and year before or in 2000 E.C, then the patient will terminate the ART care else will be on ART treatment.

If ART stage is OA and Functional status is Bedridden and the year in which the individual patient started the service is after 1999 E.C and Hospital is Asella and then the patient will terminate the ART care.

If ART stage is OA and Functional is Bedridden and the year in which the individual patient started the service is after 1999 E.C and Hospital is Adama and then the patient will not terminate the ART care.

There is some hidden pattern within the classification results that there is a shorter time attrition rate at Asella hospital than Adama hospital; i.e. the patient stays longer time at Adama hospital than in Asella hospital. This might trigger the service givers to investigate and bring necessary conducive environment to maintain the patients in ART care. They year element also came to light that those who joined the service before three years are highly likely to terminate the care.

If ART stage of the patients is IN (in care), ER (Eligible and Ready) or EL (Eligible) then they are highly likely on ART care (will not terminate). Those patients whose ART Stage is OA and Functional Status is P (appropriate) or D (Delay) or R (Regression) Then the patients has high chance not to terminate the treatment.

If ART stage of the patient is ‘OA’ and Functional status is ‘A’ and the patients are less then 18 years then the patient will not terminate the care. This hidden knowledge is predicting that youth patients are high chance to stay on the ART care for longer years.

Scenario #6 model generated tree of less complexity with little pattern which are meaningfully interpreted compared to its pruned counterpart. It is possible to say generalized decision tree with pruning with all attribute is compact and interpretable tree structure. The knowledge obtained from the pruned generalized decision tree with all attributes model has got meaningful contributions to the interventions of ART care service delivery as also supported by the opinion of the expert in the domain area. The case in mention here is the knowledge which signaled “patients who began the Antiretroviral (ARV) drug and whose health condition or functional status were bedridden high chance of terminating the care/treatment”. Fact can make the care givers to focus their attentions to maintain those with high risk of termination.

CHAPTER SIX

EXPERIMENTS AND ANALYSIS OF ASSOCIATION

The second class of data mining to be experimented is association rule mining. The 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 5.1 shows the number of experiments to be carried out with their parameters.

Table 6.1: Experiments made for Association Rule Mining

Experiment #	Minimum Thresholds	
	Support	Confidence
1	0.3	50%
2	0.25	60%
3	0.25	70%
4	0.25	80%
5	0.2	80%
6	0.2	85%
7	0.2	90%
8	0.15	85%
9	0.15	90%
10	0.15	95%
11	0.1	90%
12	0.1	95%
13	0.1	100%

So, there will be 13 scenarios/experiments at the end of the association rule mining using apriori algorithm in weka as mentioned in Chapter 4 section 4.3. The number of rules generated for those experiments are also going to be analyzed, important patterns will be obtained after sorting them with their confidence. Let as see these experiments one by one.

Experiment #1: Minimum support of 0.3 and Minimum Confidence of 50%

The model obtained from this model looks like as follows:

```
Apriori
=======

Minimum support: 0.5 (9370 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 6

Best rules found:

 1. Year='(1997-inf)' FunctionalStatus=W 13062 ==> Termination=FALSE 10222
        conf :( 0.78)
 2. Functional Status=W 13585 ==> Termination=FALSE 10546    conf :( 0.78)
 3. Year='(1997-inf)' OACD4='(96.25-191.5]' 14546 ==> Termination=FALSE 11206
        conf :( 0.77)
 4. Year='(1997-inf)' OAWHO=Stage3 13744 ==> Termination=FALSE 10524    conf
        :( 0.77)
 5. OACD4='(96.25-191.5]' 14914 ==> Termination=FALSE 11411    conf :( 0.77)
 6. OAWHO=Stage3 14226 ==> Termination=FALSE 10793    conf :( 0.76)
 7. Hospital=Adama Year='(1997-inf)' 12808 ==> Termination=FALSE 9541    conf :(
        0.74)
 8. Year='(1997-inf)' 18062 ==> Termination=FALSE 13448    conf :( 0.74)
 9. Year='(1997-inf)' Religion=Orthodox 14585 ==> Termination=FALSE 10836
        conf :( 0.74)
10. Year='(1997-inf)' Occupation=UnEmp 12743 ==> Termination=FALSE 9449
        conf :( 0.74)
```

As it can be seen from the association rule model result above, all the 10 best rules were produced at minimum support of 0.3 and minimum confidence of 50% and revealed important patterns in explaining the association that exists between year, functional status, WHO stage, OACD4, Occupation and Religion as antecedent and Termination status as Consequent.

Experiment #2: With minimum support of 0.25 and minimum confidence of 60%

The second scenario provided the top ten rules obtained. The result of this experiment is presented ad follows. Again, from the second experiment we found all the top ten rules with new as well as duplicate rules with a bit lower confidence for them than before. We will try to sort theses rules by their confidence and take the rule with greater confidence

```
        Apriori
=======

Minimum support: 0.5 (9370 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 6

Best rules found:

 1. Year='(1997-inf)' FunctionalStatus=W 13062 ==> Termination=FALSE 10222
          conf :( 0.78)
 2. FunctionalStatus=W 13585 ==> Termination=FALSE 10546    conf :( 0.78)
 3. Year='(1997-inf)' OACD4='(96.25-191.5]' 14546 ==> Termination=FALSE 11206
          conf :( 0.77)
 4. Year='(1997-inf)' OAWHO=Stage3 13744 ==> Termination=FALSE 10524    conf
          :( 0.77)
 5. OACD4='(96.25-191.5]' 14914 ==> Termination=FALSE 11411    conf :( 0.77)
 6. OAWHO=Stage3 14226 ==> Termination=FALSE 10793    conf :( 0.76)
 7. Hospital=Adama Year='(1997-inf)' 12808 ==> Termination=FALSE 9541    conf :(
            0.74)
 8. Year='(1997-inf)' 18062 ==> Termination=FALSE 13448    conf :( 0.74)
 9. Year='(1997-inf)' Religion=Orthodox 14585 ==> Termination=FALSE 10836
            conf :( 0.74)
10. Year='(1997-inf)' Occupation=UnEmp 12743 ==> Termination=FALSE 9449
            conf :( 0.74)
```

later when the all experimentation is over. The meaning of the rules will be given for those selected rules at the end of the experimentation.

Experiment#3: With 0.25 Minimum Support and 70% Minimum Confidence

```
Apriori
=======

Minimum support: 0.5 (9370 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 6

Best rules found:

 1. Year='(1997-inf)' FunctinalStatus=W 13062 ==> Termination=FALSE 10222
          conf :( 0.78)
 2. FunctinalStatus=W 13585 ==> Termination=FALSE 10546    conf :( 0.78)
 3. Year='(1997-inf)' OACD4='(96.25-191.5]' 14546 ==> Termination=FALSE 11206
          conf :( 0.77)
 4. Year='(1997-inf)' OAWHO=Stage3 13744 ==> Termination=FALSE 10524
          conf :( 0.77)
 5. OACD4='(96.25-191.5]' 14914 ==> Termination=FALSE 11411    conf :( 0.77)
 6. OAWHO=Stage3 14226 ==> Termination=FALSE 10793    conf :( 0.76)
 7. Hospital=Adama Year='(1997-inf)' 12808 ==> Termination=FALSE 9541    conf :(
          0.74)
 8. Year='(1997-inf)' 18062 ==> Termination=FALSE 13448    conf :( 0.74)
 9. Year='(1997-inf)' Religion=Orthodox 14585 ==> Termination=FALSE 10836
          conf :( 0.74)
10. Year='(1997-inf)' Occupation=UnEmp 12743 ==> Termination=FALSE 9449
          conf :( 0.74)
```

At the minimum Support of 0.3 and Confidence of 70% thresholds, the ten best rules obtained with 50% support and 74% to 78% confidence. This means that the association between the antecedent and the consequent is more sufficient than the supplied minimum thresholds. The support is the proportion of number of times the antecedent and consequent occurred together for example 10836 for rule #10 to the total number of the dataset. Confidence is the proportion of Number on the Right Hand Side (RHS), which shows the number of times antecedent and Consequent occur together in the association

rule to the number of Left Hand Side (LHS) or Antecedent. From this model it is possible to see rules like:

Rule #9: Year='(1997-inf)' Religion=Orthodox 14585 ==> Termination=FALSE 10836 conf :(0.74)

Meaning: If Year in which the patient is registered for the ART care is after 1997 and patient's Religion is Orthodox then the patient will not terminate the Care with confidence of 74% and Support of 50%.

Rule #10: Year='(1997-inf)' Occupation=UnEmp 12743 ==> Termination=FALSE 9449 conf :(0.74)

Meaning: If Year in which the patient is registered for the ART care is after 1997 E.C and the patients are unemployed then Patients will not terminate the ART Care with confidence of 74% and Support of 50%.

Experiment#4: With minimum support of 0.25 and minimum confidence of 80%

In this model the minimum support threshold doesn't change i.e. it was 25% for the generated Rules. However, the confidence provided by the model is greater than the supplies minimum confidence threshold, which is 86% for all top best rules generated. This means the proportion of the number of occurrence of Antecedent and consequent together to the number of occurrence of Antecedent is higher than the given minimum threshold. The resulting model of this scenario is presented bellow:

Minimum support: 0.25 (4685 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 53

Size of set of large itemsets L(3): 66

Size of set of large itemsets L(4): 38

Size of set of large itemsets L(5): 6

Best rules found:

1. ARTStage=IN Year='(1997-inf)' OACD4='(96.25-191.5]' 6964 ==> Termination=FALSE 6005 conf :(0.86)
2. ARTStage=IN Year='(1997-inf)' 7009 ==> Termination=FALSE 6043 conf :(0.86)
3. ARTStage=IN OACD4='(96.25-191.5]' 6992 ==> Termination=FALSE 6026 conf :(0.86)
4. ARTStage=IN 7037 ==> Termination=FALSE 6064 conf :(0.86)
5. ARTStage=IN Year='(1997-inf)' Religion=Orthodox OACD4='(96.25-191.5]' 5773 ==> Termination=FALSE 4961 conf :(0.86)
6. ARTStage=IN Year='(1997-inf)' Religion=Orthodox 5804 ==> Termination=FALSE 4987 conf :(0.86)
7. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 6513 ==> Termination=FALSE 5595 conf :(0.86)
8. ARTStage=IN Religion=Orthodox OACD4='(96.25-191.5]' 5796 ==> Termination=FALSE 4978 conf :(0.86)
9. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 OACD4='(96.25-191.5]' 6489 ==> Termination=FALSE 5573 conf :(0.86)
10. ARTStage=IN Religion=Orthodox 5827 ==> Termination=FALSE 5004 conf :(0.86)

A bit new rules with high confidence with new attribute values are obtained in this scenario.The meaning of rule #3 above is IF ART Stage is ‘IN’ (in care) and ‘OACD4’ (CD4 count of the patient) is between 96 to 192 THEN ‘termination’ (termination status of the patient) will be False (the patient is on care) with 25% support and 86%

confidence. The support of the association rule is simply the percentage of both itemsets occurred together divided by the total number of records used for building the association rule. Confidence is the percentage of the number appeared on the Consequent (RHS), which is 6026 divided by the number appeared on the antecedent (LHS), which is 6992. These issues were discussed in section 3.4.2 well thoroughly when we discuss about model and algorithms selection. This rule is a three itemset rule.

Experiment#5: With Minimum support of 0.2 and minimum confidence of 80%

The result of this experiment is as present bellow.

Minimum support: 0.25 (4685 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 53

Size of set of large itemsets L(3): 66

Size of set of large itemsets L(4): 38

Size of set of large itemsets L(5): 6

Best rules found:

1. ARTStage=IN Year='(1997-inf)' OACD4='(96.25-191.5]' 6964 ==> Termination=FALSE 6005 conf :(0.86)
2. ARTStage=IN Year='(1997-inf)' 7009 ==> Termination=FALSE 6043 conf :(0.86)
3. ARTStage=IN OACD4='(96.25-191.5]' 6992 ==> Termination=FALSE 6026 conf :(0.86)
4. ARTStage=IN 7037 ==> Termination=FALSE 6064 conf :(0.86)
5. ARTStage=IN Year='(1997-inf)' Religion=Orthodox OACD4='(96.25-191.5]' 5773 ==> Termination=FALSE 4961 conf :(0.86)
6. ARTStage=IN Year='(1997-inf)' Religion=Orthodox 5804 ==> Termination=FALSE 4987 conf :(0.86)
7. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 6513 ==> Termination=FALSE 5595 conf :(0.86)
8. ARTStage=IN Religion=Orthodox OACD4='(96.25-191.5]' 5796 ==> Termination=FALSE 4978 conf :(0.86)
9. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 OACD4='(96.25-191.5]' 6489 ==> Termination=FALSE 5573 conf :(0.86)
10. ARTStage=IN Religion=Orthodox 5827 ==> Termination=FALSE 5004 conf :(0.86)

There were a total of 12 rules out of which the top ten of them are presented above in this scenario.

Experiment# 6: Minimum support of 0.2 and Minimum Confidence of 85%

Minimum support: 0.25 (4685 instances)
Minimum metric <confidence>: 0.85
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 53

Size of set of large itemsets L(3): 66

Size of set of large itemsets L(4): 38

Size of set of large itemsets L(5): 6

Best rules found:

1. ARTStage=IN Year='(1997-inf)' OACD4='(96.25-191.5]' 6964 ==> Termination=FALSE 6005 conf :(0.86)
2. ARTStage=IN Year='(1997-inf)' 7009 ==> Termination=FALSE 6043 conf :(0.86)
3. ARTStage=IN OACD4='(96.25-191.5]' 6992 ==> Termination=FALSE 6026 conf :(0.86)
4. ARTStage=IN 7037 ==> Termination=FALSE 6064 conf :(0.86)
5. ARTStage=IN Year='(1997-inf)' Religion=Orthodox OACD4='(96.25-191.5]' 5773 ==> Termination=FALSE 4961 conf :(0.86)
6. ARTStage=IN Year='(1997-inf)' Religion=Orthodox 5804 ==> Termination=FALSE 4987 conf :(0.86)
7. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 6513 ==> Termination=FALSE 5595 conf :(0.86)
8. ARTStage=IN Religion=Orthodox OACD4='(96.25-191.5]' 5796 ==> Termination=FALSE 4978 conf :(0.86)
9. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 OACD4='(96.25-191.5]' 6489 ==> Termination=FALSE 5573 conf :(0.86)
10. ARTStage=IN Religion=Orthodox 5827 ==> Termination=FALSE 5004 conf :(0.86)

In the sixth scenario a total of 16 rules were obtained and the top ten rules are as presented above. There are also new patterns with higher confidence.

This model encompasses the patterns we have so far with the greater confidence. It has generated also the top 10 rules with better confidence.

The meaning of this rule number 3 above is ART Stage is 'IN' (in care) and if OACD4 is between 96 and 192 THEN 'termination' (termination status of the patient is) is False (will not terminate the care) with 25% support and 86% confidence.

Experiment #7: With Minimum Support of 0.2 and minimum confidence of 90%

Minimum support: 0.2 (3748 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 19

Size of set of large itemsets L(2): 73

Size of set of large itemsets L(3): 128

Size of set of large itemsets L(4): 96

Size of set of large itemsets L(5): 29

Size of set of large itemsets L(6): 2

As we can see from the result of the model, there is no single best rule found in scenario at minimum confidence of 90% for the top ten rules required.

Experiment #8: With Minimum Support of 0.15 and minimum Confidence 85%

Minimum support: 0.25 (4685 instances)
Minimum metric <confidence>: 0.85
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L (1): 13

Size of set of large itemsets L (2): 53

Size of set of large itemsets L (3): 66

Size of set of large itemsets L (4): 38

Size of set of large itemsets L (5): 6

Best rules found:

1. ARTStage=IN Year='(1997-inf)' OACD4='(96.25-191.5]' 6964 ==> Termination=FALSE 6005 conf :(0.86)
2. ARTStage=IN Year='(1997-inf)' 7009 ==> Termination=FALSE 6043 conf :(0.86)
3. ARTStage=IN OACD4='(96.25-191.5]' 6992 ==> Termination=FALSE 6026 conf :(0.86)
4. ARTStage=IN 7037 ==> Termination=FALSE 6064 conf :(0.86)
5. ARTStage=IN Year='(1997-inf)' Religion=Orthodox OACD4='(96.25-191.5]' 5773 ==> Termination=FALSE 4961 conf :(0.86)
6. ARTStage=IN Year='(1997-inf)' Religion=Orthodox 5804 ==> Termination=FALSE 4987 conf :(0.86)
7. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 6513 ==> Termination=FALSE 5595 conf :(0.86)
8. ARTStage=IN Religion=Orthodox OACD4='(96.25-191.5]' 5796 ==> Termination=FALSE 4978 conf :(0.86)
9. ARTStage=IN Year='(1997-inf)' OAWHO=Stage3 OACD4='(96.25-191.5]' 6489 ==> Termination=FALSE 5573 conf :(0.86)
10. ARTStage=IN Religion=Orthodox 5827 ==> Termination=FALSE 5004 conf :(0.86)

Rule number #3 above has a meaning of if ART Stage IN (in care) and CD4 count of the patient is between 96 and 192 then the patient will not terminate the ART service. In this scenario, there were a total of 44 rules with confidence of 86%. As you can see both the

support and confidence of the model is higher than the supplied minimum thresholds

Experiment#9: Minimum support of 0.15 and minimum Confidence of 90%

```
Minimum support: 0.15 (2811 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Best rules found:
```

This scenario revealed limited the number of association rule, to the extent of not having a single best rule out of the top 10 rules required. As observed repeatedly, at the minimum confidence of 90% threshold and minimum support threshold greater than 10%, there is no best rule found from the top 10 best rules required.

Experiment #10: with minimum support of 0.15 and minimum confidence of 95

If by implication, it is possible to infer that if we can't get any best rule at minimum support of 0.15 and confidence of 90% that there will not be any best rule by keeping minimum support the same and increasing the minimum confidence i.e. 95%. The model has also shown the same as bellow.

```
Minimum support: 0.15 (2811 instances)
Minimum metric <confidence>: 0.95
Number of cycles performed: 17

Best rules found:
```

Experiment #11: With Minimum Support of 0.1 and Minimum Confidence 90%

Minimum support: 0.1 (1874 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 32

Size of set of large itemsets L(2): 172

Size of set of large itemsets L(3): 430

Size of set of large itemsets L(4): 550

Size of set of large itemsets L(5): 412

Size of set of large itemsets L(6): 164

Size of set of large itemsets L(7): 26

Best rules found:

1. Hospital=Asella ARTStage=IN OACD4='(96.25-191.5]' 2126 ==> Termination=FALSE 1942 conf :(0.91)
2. Hospital=Asella ARTStage=IN Year='(1997-inf)' OACD4='(96.25-191.5]' 2126 ==> Termination=FALSE 1942 conf :(0.91)
3. Hospital=Asella ARTStage=IN 2140 ==> Termination=FALSE 1954 conf :(0.91)
4. Hospital=Asella ARTStage=IN Year='(1997-inf)' 2140 ==> Termination=FALSE 1954 conf :(0.91)

As you can from the above model, from the top ten rules required at minimum support of 10 % and minimum confidence of 90%, there are only 4 rules produced. The mare reason that the four rules obtained is because the minimum support supplied by the researcher is decreased by 0.05 than experiment by keeping the minimum confidence threshold the same to that experiment #10.The confidence reported by the model in this model is 91% for all the four rules generated and most of the rules which were generated by the previous models tends to be subset of this model.

Experiment #12: with minimum support of 0.1 and minimum confidence of 95%

In this experiment there is no best rule found in the effort mode to see 10 top ten rules. Increasing the minimum confidence threshold by 0.05, keeping the minimum support threshold the same to the experiment #11 results in not getting any best rule in this model The result of the model looks like:

```
Minimum support: 0.1 (1874 instances)
Minimum metric <confidence>: 0.95
Number of cycles performed: 18

Best rules found:
```

Experiment #13: With Minimum support of 0.1 and minimum 100%

There was no rule found for this experiment as you can see from the model result

```
Apriori
=======

Minimum support: 0.1 (1874 instances)
Minimum metric <confidence>: 1
Number of cycles performed: 18


Best rules found:
```

Summary of number of rules obtained from the above experiments is presented in Table 6.2

Table 6.2: Number of rules each minimum Thresholds

Experiment #	Minimum Thresholds		#Rules Of 100 required
	Support	Confidence	
1	0.3	50%	100
2	0.25	60%	53
3	0.25	70%	92
4	0.25	80%	12
5	0.2	80%	30
6	0.2	85%	0
7	0.2	90%	0
8	0.15	85%	44
9	0.15	90%	0
10	0.15	95%	0
11	0.1	90%	4
12	0.1	95%	0
13	0.1	100%	0

The number of rule is decreased when the minimum support and minimum confidence is increasing. Keeping minimum support constant and increasing minimum confidence

threshold again decreases the number of rules. Increasing minimum support; keeping the minimum confidence again reduces the number of rule generated.

The following rules were selected from different experiments for interpretation collecting, comparing them for new itemset and sorting them by their confidence

Rule#1: Hospital=Asella ARTStage=IN OACD4='(96.25-191.5]' ==> Termination=FALSE

Meaning: IF Hospital in which the patient is treated is Asella and the patients ART stage is in care then the patient will not terminate the Treatement

Performance: (10% Support, 91% Confidence**)**

Rule#2: Hospital=Asella ARTStage=IN Year='(1997-inf)' OACD4='(96.25-191.5]' ==> Termination=FALSE

Meaning: If Hospital is Asella and ART stage is in care and year is after 1997 and CD4 count is between 96 and 196 then termination status is False.

Performance: (10% support and 91% Confidence)

Rule#3: ARTStage=IN FunctionalStatus=W OACD4='(96.25-191.5]' ==> Termination= FALSE conf :(0.86)

Meaning: IF ART stage is "IN" (in care) and Functional Status is "W" (working) and "OACD4" (CD4 count) is between 96 and 192 then "Termination" Status is False.

Rule #5: ARTStage=IN Year='(1997-inf)' Religion=Orthodox OACD4='(96.25-191.5]' ==> Termination=FALSE

Meaning: If ART Stage is in care and Year in which the patient is started the service is after 1997 and religion is orthodox and OACD4 is between 96 and 192 then the termination Status is false

Performance: (20% Support, 86% confidence)

If ART Stage of the patients is 'IN' (in care) and their Functional is 'W' (Working) Then they will have termination Status "False".

Being on 'IN' ART stage is associated to false value of 'Termination'.

If ART Stage is 'IN' and Functional Status is 'W' and WHO clinical stage of patient is

stage 3 and CD4 count is between 96 and 1992 Then with 86% confidence the patient will not terminate the ART care service.:

Most of the patterns were repeated in all of the three scenarios unless the difference in the number of rules generated, support, and confidence difference. Most of patterns were also found in the decision tree models. The minimum experiments done on both decision tree models and association rule model with different parameters, the researcher discovered data mining has an applicability to discover hidden and none trivial patterns or relationships within the dataset since they met the minimum thresholds in both cases as shown in the above analysis . The patterns were also appreciated by the domain experts working on the area.

CHAPTER SEVEN

CONCLUSIONS AND RECOMMENDATIONS

In this chapter concluding remarks regarding the research processes, experimentations made and attained results as well as important recommendations are made.

6.1 Conclusions

Health care in general is data reach and knowledge poor as stated. Clinical services delivered at facility level like in hospitals takes tremendous demographic as well as clinical history of the patients who are treated at the facilities. This data grows exponentially from time to time requiring powerful analysis tools for producing decision support information. Data miming techniques are solutions in discovering non-trivial, hidden, and potentially useful decision support information/knowledge out large volume of data collected overtime from many sources.

Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) pandemic continues to spread worldwide. Very large number of people (range 34.6-42.3 million) were living with the virus, which killed about 3 million in 2003, and over 20 million since the first case of AIDS were identified in 1981. The situation is more exacerbated in Sub Saharan African region which carries 80% the world HIV/AIDS burden accounted for 10% of the total number the world population. Ethiopia is one of the Sub-Saharan Africa which is also badly affected by the disease having more than 1 million people living with HIV/AIDS. The catastrophic spread of the disease urged the world to work hard to curb the problem. The only best achievement to date is the ability of enabling individuals infected with HIV to lead disease free life through Antiretroviral Therapy (ART).

The therapy came to those who are in need at large worldwide freely starting from 2005 onwards. Ethiopia has also aggressively worked on bringing HIV infected people to the ART care since then. However the service delivery is exacerbated by many factors which

need further research. On the other hand; a lot of data is collected from patients registered for the service in addition to the increasing number of patients coming to the services. In Ethiopia, 364,724 people were on pre ART and ART out of which 183,583 are on ART care in 2008. Therefore; it is clear that the data about these patients collected at the facilities can reveal some hidden but important information for decision making regarding maintaining those patients in the care. In discovering these hidden patterns; data mining has high potential. This was the intention of this research i.e. assessing applicability of data mining techniques in finding out patterns which affect termination/ continuity behavior of HIV infected people once they are registered for the ART Care service.

For conducting the research, data was collected from the two hospitals (Adama and Asella hospitals) taken as cases for investigation. Data preprocessing in preparing the data for model building was conducted by the researcher. Attribute selections were made. Models were selected for data mining tasks based on its appropriateness for the problem at. Data preprocessing lead to more concise and understandable knowledge after model building. The modeling methodology applied to the research was CIRSP-DM (Cross Industry Standard for Data Mining), which involves business understanding, data understanding, data preparation, model building, evaluation of the models and deployment steps/phases. Different literatures were reviewed with effort to bring the problem into data mining problem

Tools used in the research include Ms-excel for preliminary data understanding and analysis purpose of results of the model, EpiInfo for preprocessing tasks, Weka for model building and some preprocessing tasks, Ms-Word for documentation purpose and power point for presentation purpose.

Selected data mining methods implemented for this research were classification and association. For classification models, decision tree were selected for easy structure and visualization power it has. Eight experiments were undergone for classification purpose. These are binary decision tree with and without pruning with all and some attributes, generalized decision tree both with pruning and without pruning with all and some

selected attributes. These experiments were performed with J48 decision tree implementation in weka. From these experiments it was found that all of the scenarios of decision tree classification are applicable on the dataset moderately but the generalized decision tree outperforms than others in terms accuracy measures based confusion matrix as well as in generating tree structures which are compact and whose patterns are easily understandable and interpretable. Size of the generalized decision tree with pruning is by far smaller and had less number of leaves than the others even though it is still impossible to visualize its tree structure by weka in a readable form. But the tree result buffer of this scenario was too compact to present it in the body of the research.

Attribute ordering algorithm of weka was also run and ranks of the attributes were known for classification model. ART stage comes first in determining continuity/termination behavior of the patients and Hospital attribute comes last. Accordingly ART stage, Functional status, Year, OACD4, OAWHO and with their values were the top six determining attributes for the termination behavior of the patients as they in appeared with meaningful pattern at whatever level of abstraction. The firs three with last attribute appeared in the generalized decision tree as well association rules of high value minimum thresholds.

It was learned that increasing minimum support threshold keeping minimum confidence at 0.9, it reduced the number of rules to extent of not having any best rule. As we decrease the confidence and increase the minimum support, more rules with meaningful pattern were obtained. As we decrease the minimum support from steadily 0.3 to 0.1 and increasing the minimum confidence from 50% to 90%, it provided encouraging patterns with acceptable performance (Support and Confidence).

Important patterns were obtained in terms of years, ART stage, Age of the patients in determining continuity/ termination behavior of patients for ART care. To mention few:

If ART Stage of the patients is ‘IN’ (in care) and their Functional is ‘W’ (Working) then they have high chance not to terminate the care. Being on ‘IN’ ART stage is associated to false value of ‘Termination’ with the confidence of. If ART Stage is ‘IN’ and Functional

Status is ‘W’ and WHO clinical stage of patient is stage 3 and CD4 count is between 96 and 1992 Then with 86% confidence the patient will not terminate the care. There was also hidden knowledge that there is a length year difference to stay on ART by the patient Those who were treated at Adama hospital had longer continuity those treated at Asella hospital. An other important knowledge is youngsters have better continuity than others.

Therefore the main aim of the research was met here with identifying the best performing scenario of data mining the techniques with knowing the most determining factor/attribute for termination behavior of the ART care by the patients on the dataset taken for the research. With this the researcher proves that applicability of the data mining techniques on the ART data on the taken cases.

In general, the results from this study were encouraging. It was possible to identify the frontier determining attributes and their values for the termination/continuity status of patients to ART care using data mining techniques that made good meanings to domain experts. The generalized decision tree with pruning found to be the top relevant technique on the dataset to get meaningful patterns from the decision tree experiments. Association rules mining was happened to be of high importance in its compactness and in bringing new attributes as a determining factors for termination status like occupation which were not presented in the generalized decision tree with pruning scenarios. The researcher’s believe that a more thorough study using data mining techniques can help to understand more about determinant factors of the ART care termination/ continuity behavior once the patients started it, a problem needs to be addressed well in Ethiopia. Thorough discussion with domain experts on the discovered patterns helps for getting meaningful decision support information for solving the continuity problem of ART service of patients.

6.2 Recommendations

Even though the research is done for academic achievements; the research output would help care givers in signaling targets of high risk not to stay in the ART care and enable to focus on solving the problems which hinders the continuity of ART care by the patients.

More coverage is needed in terms of facilities from which the data is taken. But the researcher believes this research can be a corner stone for the research works done in the area of ART dataset in future.

Database standardization is needed for the databases from which the data is taken as there were so many invalid values in many of the attributes which consumes much of the researcher time to clean.

The researcher experimented only two classes of data mining techniques; again only limited number of the scenarios were tested but those data mining techniques which were not experimented by the researcher might reveal important patterns in relation to factors affecting continuity/termination behaviors of the patients on ART; therefore it is left open for future research works to explore this problem in future.

A thorough discussion with the domain experts is needed on the patterns discovered to identify the most meaningful decision support information to deploy it to the benefit of the patients and hence the successfulness of ART Program.

Lastly not least; the problem of getting health related data for data mining researches is immense even though the ethical considerations are made; this problem should be improved for health informatics researchers in the future.

References

Abdul-Kareem, S.Et. Al. 2000. ANN as a Tool for Medical Prognosis. Sagepub

Abraham, Tesso.2005. Application of Data Mining Technology to Identify Determinant Risk Factors Of HIV Infection To Find Their Association Rules: The Case of Center For Disease Controls And Prevention (CDC).Unpublished Thesis, Addis Ababa University, Department Of Information Science.

Agrawal, R. and Srikant R.1994. Fast Algorithms for Mining Association Rules. IBM Almaden Research Center

Bach, M. And Ćosić, D.2007. Data Mining Usage in Health Care Management. Sveučilište U Zagrebu

Balac,N., 2006.Data Mining For Scientific Applications. Educational Training Workshops proceeding.US

Baylis P. 2007. Better Health Care With Data Mining. Shared Medical Systems Limited, UK :

Berry, M.,&Linoff, S. (2000). Mastering Data Mining: The Art and Science of Customer Relationship Management. New York:Wiley.

Bouckaert, R. et al.2008.Weka Manual For Version 3-6-0. University of waikato

Bramer, M. 2007. Principles of Data Mining. Springer. London Limit

Chapman, P. et. al .2000.CRISP-DM 1.0: Step-By-Step Data Mining Guide. SPSS

Ethiopian Ministry Of Health.2004. Strategic Plan for Multi Sectoral Response against HIV/AIDS For 2004-2008.Addis Ababa, Ethiopia, Ministry Of Health.

Ethiopian Ministry Of Health .2007. Antiretroviral Treatment Program Implementation Guideline In Ethiopia. Addis Ababa (Ethiopia): Ministry Of Health.

Ethiopian Ministry Of Health Disease Control And Prevenention Department Report.2004.AIDS In Ethiopia. Ministry of Health of Ethiopia, Addis Ababa

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

Grossman, R .1998.Data Mining Research: Opportunities and Challenges. Epapers.Dmr-V8-4-5.

HAPCO Report .2008. UNAIDS. Addis Ababa, Ethiopia

Health Services Research and the HIPAA Privacy Rule. Available at: URL <Http://Privacyruleandresearch.Nih.Gov/Healthservicesprivacy.As>

Houston, A. et al.2000. Medical Data Mining on the Internet: Research on a cancer information system: Artificial Intelligence Review 13: 437–466.

IRM .1999. Data Mining Discovering Opportunities in Your Company Data. Available at URL: http://www.neural.uom.gr/documents /data% 20mining/ data%20Mining.pdf on 16/11/2009 at 3: 00 PM

UNAIDS.2007. Uniting The World Against AIDS Available at :Http://www.unaids .org /En/ Knowledgecentre/Hivdata/ 2007

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

Kaur,H. And Wasan, S. 2000. Empirical Study on Applications of Data Mining Techniques in Healthcare, New Delhi.

Kruse G, Et Al. (2009) Burnout and use of HIV services among Health Care Workers in Lusaka District, Zambia: A Cross-Sectional Study. Biomed Central Ltd; Lusaka, Zambia.

Levin, Nissan And Zahavi, Jacob, 1999. Data Mining. Available URL:Www. Urbanscience.Com/Data_Mining.Pdf

Mannila, Heikki. 2002. Methods And Problems In Data Mining. Available URL: Http: //www.cs.helsinki.fi/~mannila/

Morris, J. Et. Al. 2007.Injection Drug Use and Patterns Of Highly Active Antiretroviral Therapy (HAART) Use: An Analysis Of ALIVE, WIHS, And MACS Cohorts. Biomed Central Ltd. AIDS Resv.4

Piatetsky-Shapiro, D.1996. Selecting And Reporting What is Interesting. in Advances in Knowledge Discovery and Data Mining. Mcneil, AAI Press,

Seifert, J.2004. Data Mining: An Overview, CRS Report For Congress.IRP

Seoulaja .2009. K-Means Clustering Algorithm Data Mining Tutorial.

Shagaw, Anagew.2002.Application Of Data Mining Technology To Predict Child Mortality Pattern: The Case Of Butajira Rural Health Project. Unpublished

Thesis, Addis Ababa University, Department Of Information Science.

The United States President's Emergency Plan for AIDS Relief .2007.Available at: http://www.pepfar.gov/. Accessed 25 March 2009

Two Crows Corporation.1999. Introduction To Data Mining And Knowledge Discovery. Twocrows

Raghavan, Vijay, Deogun, Jitender S. and Sover Mayri, 2002. Data Mining : Trends and Issues.

UNAIDS 2004 Report On The Global AIDS Epidemic - 4th Global Report on AIDS Epidemic. Bangkok

Vassilis.2008.Data Mining Techniques for HIV/AIDS Data Management in Thailand. Journal of Enterprise Information Management, 21 (1). Pp. 52-70

WHO.2008. Towards Universal Access Scaling up Priority HIV/AIDS Interventions in The Health Sector. EMH

Witten IH, Frank E. Data Mining: Practical Machine Learning Tools And Techniques. Second Edition, 2005. Morgan Kaufmann.