/


ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

# APPLICATION OF DATA MINING FOR PREDICTING ADULT MORTALITY

TESFAHUN HAILEMARIAM

JUNE 2012

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

# APPLICATION OF DATA MINING FOR PREDICTING ADULT MORTALITY

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

By

TESFAHUN HAILEMARIAM

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

# APPLICATION OF DATA MINING FOR PREDICTING ADULT MORTALITY

By

TESFAHUN HAILEMARIAM

Name and signature of Members of the Examining Board

| **Name** | **Title** | **Signature** | **Date** |
|---|---|---|---|
| _____ | Chair person | _____ | _____ |
| _____ | Advisor(s**),** | _____ | _____ |
| _____ | Advisor(s), | _____ | _____ |
| _____ | Examiner**,** | _____ | _____ |
| _____ | Examiner, | _____ | _____ |

**DECLARATION**

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

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

_____

Advisor

# DEDICATION

This study is dedicated to those who are working on adult mortality reduction to wipe away tears of family and to ensure economic ramification of the country.

# ACKNOWLEDGEMENTS

It is a great pleasure for me to express my heartfelt gratitude to Dr. Million Meshesha, who has been an excellent advisor and mentor. Million has taught me a great deal about data mining researches and guide me in a proper direction to achieve the objective of this research. I am also grateful to my advisor Dr. Alemayehu Worku for his bringing the research area to my attention, keen insight, and for his all rounded guidance and support.

I will extend my thanks to the Addis Ababa University for sponsoring this thesis Research.
My appreciations also go to all staffs of Information Sciences and Public Health Department.

Helpful hands have been extended throughout the whole process, Abebe Lolamo, Beemnet Tekabe, Minale Tefera, and my classmates need to be recognized and duly acknowledged.

Special thanks go to my parents, brothers, and sisters for their unreserved all rounded support, undying prayers, and love they have for me. 'Thank you'

Above all, I praise the Almighty God, for the strength and endurance He gave me
all the time. Glory be to him!!!

# TABLE OF CONTENTS

ACKNOWLEDGMENTS.....i
LISTS OF TABLES..... v
LISTS OF FIGURES.....vi
ACRONMYS AND ABBREVIATIONS.....vii
ABSTRACT .....viii
CHAPTER ONE.....1
INTRODUCTION .....1
1.1 Background.....1
1.1.1 Health care practice and Butajira rural health programme .....1
1.1.2 Data mining and the health care.....2
1.1.3. Importance and challenges of data mining.....4
1.1.4 Adult mortality conditions .....6
1.2 Statement of the Problem.....7
1.3 Objective of the Study .....10
1.3.1 General objective .....10
1.3.2 Specific objectives .....10
1.4 Scope and Limitation of the Study.....10
1.5 Research Methodology .....11
1.6 Significance of the study.....15
1.7 Ethical Considerations .....15
1.8 Organization of the Thesis .....16
CHAPTER TWO .....17
LITERATURE REVIEW .....17
2.1 Overview of Data Mining .....17
2.2 Methodology of Data Mining Research.....18
2.3 Data Mining Function.....26
2.3.1 Classification.....27
2.3.2 Clustering.....31
2.3.3 Association rule discovery.....32

2.4 Related Works ..... 35
CHAPTER THREE ..... 39
TECHINIQUES FOR MINING BRHP DATA ..... 39
3.1 Methods of Problem Domain Understanding ..... 40
3.2 Methods of Data Understanding and Data Preparation ..... 40
3.3 Methods of Modeling ..... 41
3.3.1 J48 Decision tree algorithm ..... 41
3.3.2 Naïve Bayes classifier ..... 43
3.4 Methods of Training and Testing ..... 45
3.5 Methods of Analysis and Evaluation of System Performance ..... 46
3.5.1 Confusion matrix ..... 46
3.5.2 Receiver Operating Characteristic (ROC) Curve ..... 48
CHAPTER FOUR ..... 50
BUSINESS UNDERSTANING AND PREPARATION OF BRHP DATA ..... 50
4.1 Problem Domain Understanding ..... 50
4.1.1 Task-1: Health promotion ..... 50
4.1.2 Task -2: Disease prevention ..... 53
4.1.3 Task-3: Advocacy efforts ..... 54
4.2. Data Understanding ..... 56
4.2.1 The raw data descriptions ..... 56
4.2.2 Attributes selection for knowledge discovery ..... 58
4.2.3 Descriptive data visualization ..... 59
4.3 Data Preprocessing for Mining ..... 62
4.3.1 Handling Missing Values ..... 63
4.3.2 Handling outliers ..... 64
4.3.3 Data decoding ..... 65
4.3.4 Attributes transformation ..... 66
4.3.5 Weka understandable format ..... 67
CHAPTER FIVE ..... 68
EXPERIMENTATION ..... 68
5.1 Overview of Experimentation ..... 68
5.2 Selecting and Evaluating the Attributes ..... 71
5.3 Model Building ..... 72

5.3.1 Selected model performance and evaluations ..... 75
5.4. Discussion ..... 78
5.4.1Rule Extraction ..... 79
5.5 Error Rate of the Selected Model ..... 82
CHAPTER SIX ..... 84
CONCLUSION AND RECOMMENDATION ..... 84
6.1 Summary and Conclusion ..... 84
6.2 Recommendations ..... 86
REFERENCES ..... 88
ANNEXES ..... 94
Annex 1: Calculation for Outlier Detection ..... 94
Annex 2: Outputs of the Classifiers in Experimentation ..... 95
Annex 3 : Description of the Selected Attributes ..... 98
Annex 4: Partial Decision Tree Generated for BRHP ..... 99

# LISTS OF TABLES

Table1.1: Summary of Phases and Tasks of Hybrid Model ..... 14
Table 2.1: Summary of data mining models ..... 25
Table 3.1:Confusion Matrix with Two Classes Classification Result ..... 46
Table 3.2:Performance Measures of ROC Area ..... 49
Table 4.1:Some of Adult Mortality Predictors during Business Review ..... 53
Table 4.2:List of Variables in the Initial Dataset ..... 57
Table 4.3:Selected Subset Attributes from BRHP ..... 60
Table 4.4:Dataset Selection Procedures ..... 62
Table 4.5:Statistical Summary of Attributes with Missing Values ..... 63
Tale 4.6:Missing Values Handling Mechanisms ..... 63
Table 4.7:Table for Twenty Fifth and Sevety Fifth Percentile ..... 64
Table 4. 8:Peasant association Attribute Transformation ..... 66
Table 4.9:Summary of Original and Target Datasets ..... 67
Table 5.1:Experiments and Scenarios ..... 69
Table 5.2:Attributes Rank with Information Gain ..... 70
Table 5.3:J48 Classifier Parameter Options ..... 71
Table 5.4:Performance of the Classifier for Different K-Values ..... 72
Table 5.5:Performance Summary of the Models ..... 73
Table 5.6:Sample of Instances that Shows Predicted and the Actual Class ..... 83

# LISTS OF FIGURES

Figure 1.1 Data Mining Model Used in this Research ..... 12
Figure 2.1 KDD process ..... 19
Figure 2.2 The CRISP-DM knowledge discovery Process Model ..... 21
Figure 2.3 SEMMA Process model ..... 23
Figure 2.5 Simple Feed Forward Neural Network ..... 30
Figure 3.1 Diagrammatic Overview of the Overall Research Design and Methodology ..... 39
Figure 3.2 Simple Decision Tree Constructed for Two Class Classification ..... 42
Figure 3.3 Examples for ROC curve ..... 48
Figure .4.1 Conceptual Framework for Factors Affecting Adult Mortality ..... 55
Figure 4.3 Box Plot to Detect Outliers ..... 64
Figure 4.4: Box Plot used for Numeric attributes to Detect Outliers ..... 65
Figure 5.1 Side by side Review of the class variable using SMOTE ..... 69
Figure 5.2 Learning Progress Curve ..... 72
Figure 5.3 Accuracy of All Models in the Experiments ..... 74
Figure 5.4 Performance Comparison of the Models ..... 74
Figure 5.5 J48 Pruned Tree Model with All Attributes ..... 76
Figure 5.6 ROC Area of J48 Pruned Model with all attributes ..... 77
Figure 5.7 Performace Measures the Selected Model ..... 78
Figure 5.8 Partial Decision Tree Generated For BRHP Dataset ..... 79

## ACRONYMS/ABBREVIATIONS

| | |
|---|---|
| **BRHP** | Butajira Rural Health Programme |
| **CLI** | Command Line Interface |
| **CRISP** | Cross-Industry Standard Process for Data Mining |
| **DM** | Data Mining |
| **DSSs** | Demographic Surveillance systems |
| **FP** | Frequent pattern |
| **HIV/AIDS** | Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome |
| **HSDP** | Health Sector Development Program |
| **ICT** | Information and Communication Technology |
| **KDD** | Knowledge Discovery in Data base |
| **KDP** | Knowledge Discovery Process |
| **MDGs** | Millennium Development Goals |
| **NN** | Neural Network |
| **ROC** | Receiver Operating Characteristics |
| **SAS** | Statistical Analysis System |
| **SEMMA** | Sample Explore Modify Model Assess |
| **SMOTE** | Synthetic Minority Oversampling Technique |
| **SNNPRG** | Southern Nations, Nationalities and Peoples Regional State |
| **SPECT** | Single Photo Emission Computed Tomography |
| **SPSS** | Statistical Package for Social Sciences |
| **VIF** | Variance Inflation Factor |
| **WEKA** | Waikato Environmental for Knowledge Analysis |

# ABSTRACT

**Background**: The fast-growing, tremendous amount of data, collected and stored in large and massive data repositories, has far exceeded human ability for comprehension without powerful tools. As a result, data collected in large data repositories become seldom visited. This in turn, calls the application of data mining technology. Every year, more than 7·7 million children die before their fifth birthday. However, over three times those of nearly 24 million adults die every year. Less attention has been given to adults which are the most productive phase of life for both economic and social ramification of families and countries.

**Objective:** The general objective of this research is to construct adult mortality predictive model using data mining techniques so as to identify and improve adult health status using BRHP open cohort database.

**Methods:** The hybrid model that was developed for academic research was followed. Dataset is preprocessed for missing values, outliers and data transformation. Decision tree and Naïve Bayes algorithms were employed to build the predictive model by using a sample dataset of 62,869 records of both alive and died adults through three experiments and six scenarios.

**Result:** In this study as compared to Bayes, the performance of J48 pruned decision tree reveals that 97.2% of accurate results are possible for developing classification rules that can be used for prediction. If no education in family and the person is living in rural highland and lowland, the probability of experiencing adult death is 98.4% and 97.4% respectively with concomitant attributes in the rule generated. The likely chance of adult to survive in completed primary school, completed secondary school, and further education is (98.9%, 99%, 100%) respectively.

**Conclusion:** The study suggests that education plays a considerable role as a root cause of adult death, followed by outmigration. Further comprehensive and extensive experimentation is needed to substantially describe the loss experiences of adult mortality in Ethiopia.



**Key words:** BRHP data**,** Mortality, Adult, predictive model, J48 decision tree, Data Mining.

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background

### 1.1.1 Health care practice and Butajira rural health programme

Living long is a much desired aspiration by everyone. This is because living is not only as a state of being itself valued, but also it is a necessary requirement for carrying our plans that we have reason to value [1]. Promoting health, preventing disease and prolonging life are important focuses of public health care practice. It is concerned with addressing threats to the overall health of a community through health care planning and intervention. In the broad field of public health, epidemiology is concerned with identification of underlying factors with the goal of improving health of the population through identifying the cause of death i.e. chain of factors that have given rise to an immediate medical conditions [2, 3].

Understanding this phenomenon has an important implication in terms of health promotion, disease prevention and treatment of illness by intervening at different levels of factors that helps to avoid unnecessary and the unfinished death among the population [3]. However, such conceptualization of promotive and preventive aspects of health care depend on population based analysis which usually needs timely, accurate, complete and adequate information on demographic characteristics and predisposing factors of health problems in the population [2].

In most developing countries including Ethiopia, because deaths are unregistered and nearly all take place outside health facilities, it is difficult to identify the causes of death among different population groups [3]. Consequently, provisions of appropriate interventions and evaluations become a difficult task [3]. This is mainly due to lack of accurate knowledge on complete vital registration systems of the event [4]. Thus, adult mortality levels and trends in the developing countries become hampered. As a result of the absence of systematically organized registration of vital events in developing countries, adequate and reliable health information is often lacking [5, 3]. Hence, health service planning and utilization become limited both at national and

regional levels. Therefore, population studies can fill this gap of data inadequacy and problem of health care utilization in developing countries. This in turn calls longitudinal population based studies to generate sound data on morbidity, mortality, and fertility through DSSs (Demographic surveillance systems) [6, 7].

DSSs have been established during the past 30 years in a number of field research sites in various parts of the developing countries where routine vital registration systems were poorly developed or nonexistent [3]. In Ethiopia, such a trends were established and data collection on vital events like death, birth, and related research has been conducted for the last 22 years (1986-2008) [3]. This is mainly to track a limited and common set of key variables that deal about population dynamics and demographic trends. Therefore, DSSs has an approach to define key variables with their relationships and a developed system for collection, storage and analysis of surveillance data [8].

BRHP (Butajira Rural Health Programme) has been established in 1986 as an epidemiological study that approaches to identify major variables and their relationship [6]. It is a set of field and computing operations to handle the longitudinal follow up of well defined entities (individuals, households, and residential units) and all related demographic and health outcomes within a clearly circumscribed geographic area [6]. Provision of up to date epidemiological information system to improve primary health care management and decision-making, particularly at district level is an ultimate aim of the program [2]. Events registered by the BRHP are birth, death, marriage, new household, out-migration, in-migration, and internal move (migration within the BRHP surveillance villages) [5, 8]. National and international publications and scientific conferences are the main routes of dissemination of information.

### 1.1.2 Data mining and the health care

The practice of using concrete data and evidence to support medical decisions has existed for centuries. For example, in 1839, William Farr took responsibility for medical statistics in the office of the registrar general for England and Wales. He extended the epidemiologic analysis of morbidity and mortality data, looking at effects of marital status, occupation, and altitude [9]. John Snow used maps with early forms of bar graphs in 1854 to discover the source of cholera

that resulted to death and prove that it was transmitted through the water supply [10]. Florence Nightingale invented polar area diagrams in 1855 to show that many army deaths could be traced to unsanitary clinical practices and said that it is preventable [10]. Farr, Snow and Nightingale were able to personally collect, sift through and analyze the mortality data during their times because the volume of the data was manageable.

Nowadays, the fast-growing, tremendous amount of data which is collected and stored in large and numerous data repositories has far exceeded human ability for comprehension without powerful tools [11]. As a result, data collected in large data repositories become seldom visited. Consequently, important decisions are often made not based on the information-rich data stored in data repositories rather on a decision makers' intuition. This is due to decision makers lack the tools to extract the valuable knowledge embedded in the vast amounts of data [11].

Though knowledge is a significant asset of any organization especially for information technology driven societies, today, the size of the population, the amount of electronic data gathered along with globalization and the spread of disease outbreaks and epidemics make it almost impossible to accomplish with traditional tools due to the size of the data appearing. As the volume of data increases, it is also difficult to use statistical tools to discover unanticipated complex relationship in real-world database [10, 11].

To explore unanticipated complex relationships in non-linear data property in which the epidemiological tools are inefficient and unable to discover new and interesting patterns, data mining has evolved as a new technique and methods to evaluate, analyze, search and discover new patterns and relationships hidden in large database [10].

This is where data mining becomes useful to health care and health related activities so as to explore hidden knowledge from the real world database that consist records with a complex relationship between/among variables [11]. Though data mining and its application to medicine and public health is a young field of study, it is becoming a popular and increasingly applied to tackle various medical problems through discovering the hidden patterns in health care industry. Data mining provides automated pattern recognition and it attempts to uncover patterns in data that are impossible to detect with traditional statistical methods.

People refer to data mining as the process of acquiring information, where as others refer data mining as utilization of statistical techniques within the knowledge discovery process but generally, data mining today is the set of procedures and techniques for discovering and describing patterns and trends in data [10]. Han and Kamber [11] mentioned that data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies and knowledge bases as well as scientific, social and medical research areas.

Data mining tasks are in general classified in to two main categories: predictive and descriptive data mining [12]. Predictive modeling is one of the data mining tasks that allow learning a mapping from an input set of vector measurements to a scalar output in which the training data consists of pairs of measurements. The goal of predictive modeling is to estimate (from the training data) a mapping or a function that can predict a value given an input vector of measured values and a set of estimated parameters for the model [10, 13]. The second category of data mining function is descriptive mining task used to characterize the general properties of the data in the database [11].

### 1.1.3. Importance and challenges of data mining

There is a vast potential for data mining applications in healthcare [10]. Such as evaluation of treatment effectiveness, management of healthcare, customer relationship management, detection of fraud and abuse are some of the health care area where data mining is needed. In public health care, data mining has a sound importance such as data overload: in this case data mining are best-suited for a wealth of knowledge in medical health records; policy-making in public health: it helps to discover patterns among health institutions that lead to policy recommendations and better performance in decision-making; early detection and prevention of diseases is another importance for instance classification algorithms of DM (Data Mining) help in the early detection of disease like heart disease and use as a tool to aid in monitoring trends in the clinical trials, non-invasive diagnosis and decision support.

With this regard, data mining aids provision of sophisticated treatment for clients through avoiding invasive, costly and painful diagnostic and laboratory procedures. Data mining

algorithms have a better predictive performance. For example, biopsy in women to detect cervical cancer; early detection, public health policy formulation and management in pandemic diseases; discovery of knowledge about drug side effects from the large medical database that helps to prevent patients from adverse effects of drugs are some of the data mining application in health care sectors [10].

Though data mining applications greatly benefiting the health care industry, there are some challenges in data mining [10]. Some of these are first, health care data mining can be limited by the accessibility of data because the raw inputs for data mining often exist in different settings and in different formats; second: data problems may arise in data mining application like missing values, noises, inconsistent or non-standardized data; another challenge is a large *datasets* with many variables will certainly yield random fluctuations i.e. many patterns and relationships that may not be useful for domain area; finally, data mining applications in health care organizations need a substantial investment of resources, particularly time, effort, and money otherwise it ends with failure.

There are research works performed in Ethiopia on the application of data mining technologies in different area. Shegaw [7] conducted application of data mining technology to predict child mortality patterns, Amanuel [14] applied data mining techniques to predict household health seeking patterns. Taddesse [15] applied data mining techniques to discover knowledge that can be used to gain insights in to vital statistic aspects. Helen [16] conducted application of data mining technology to identify significant patterns in census or survey data, the case of child labor survey Federal Democratic Republic of Ethiopia Central Statistics Authority. Although mortality was investigated in earlier studies, exploration on adult predictive model has not yet been attempted. In this thesis, we explored data mining technology to build a model that extracts the mortality experience of adults through middle age (15-60) over a period of 22years.

[1] *Dataset is a file of related records held on computer*.

### 1.1.4 Adult mortality conditions

Adult mortality is the probability that a 15 years old person will die before reaching his/her 60th birthday or probability of dying between 15 to 60 years per 1000 population [4, 17]. Mortality is considered as the most basic health outcome indicator [3]. In particular, age specific mortality is a widely used indicator to measure health status of the population and used to compare mortality across population groups. In a given population, adults comprise the great majority of the labor force, and it is to be expected that adult ill health and death have a deleterious effects on the productivity and well being of the population groups since adult women and men are considered as care providers of both family and community [3].

Every year, more than 7·7 million children die before their fifth birthday; however, over three times of 7.7 million of adults i.e. nearly 24 million die under the age of 70 years [18]. The risk of a 15-year-old dying before reaching 60 years of age is 12% for men and 5% for women in developed countries where as the risk of dying is double in developing countries which is 25% and 22% for men and women respectively [3].

In Ethiopia, despite a major progresses that have been made to improve the health status of the population for the last one and half decades, people still facing a high rate of morbidity and mortality and the health status of population is remained poor [21]. The pyramidal age structure of the population has remained predominately young i.e. over half (52%) of the population is in the age group of 15 and 65 years [21].

Though adults are care providers and risk takers of a society, reports indicate that adult mortality condition are not given much emphasis. This is due to a widespread perception that mortality among adults is low [3]. The increase in adult population and related changes in population health that follow demographic shift should note attention for adult health. However, there has been much less global health focus on the health and survival of adults; this in turn calls planners and policy makers to do advocacy efforts on prevention of premature adult death and routine monitoring of adult mortality [20].

Interest in adult mortality has been intensified through the MDG (Millennium Development Goals). MDG5 declaration on maternal health focuses on one of the important causes of death in women aged 15–49 years to track its goal. In order to attain the millennium promise regarding to adult health particularly, maternal mortality ratio, adult female mortality rates are pledged as an essential component of the measurement [18].

Recognizing adult death as a crucial concern for global health, nations at the millennium summit adopted the millennium declaration concerning on causes of adult death which are important components of MDGs 5 and 6. Pledges of emphasis on adult health is due to adult mortality has received little policy attention, resources and monitoring efforts [18].

The HSDPIV (Health Sector Development Program) of Ethiopia states that preventive care for adults is a key intervention and strategy to health system [21]. Therefore, monitoring progress towards achievement of MDGs that targets combating disease in promotion of health is crucial. Especially, creating a predictive model for adult mortality pattern with associated predictors is vital in order to hasten adult health within the context of the national pledges and to inform planning, programming and to guide advocacy efforts on adult health.

Thus, applying the data mining techniques is intended to address multifarious problem associated with adult health and to extract useful knowledge from the Epidemiological database of BRHP. Exploring data mining technology to predict the risk of adult mortality based up on community based epidemiological datasets gathered by the BRHP study is considered as the main task to build predictive model after identifying socio-demographic and other relevant predictors that are associated with adult mortality.

## 1.2 Statement of the Problem

The disease burden from communicable and non-communicable diseases among adult is rapidly increasing in developing countries due to ageing, health transitions and some other predictors of adult death [4, 17]. With a life expectancy of only 46 years, Ethiopia unfortunately is one of the countries with highest adult mortality rates in the world [22].

The reason for reversal in the adult life expectancy in Africa and some of the developing nations is due to premature adult deaths and the wide gap of ranges in developed and developing societies [23]. Death among adult accounts almost 12 fold mortality gaps between developed world and developing world [23]. These gaps possibly can be narrowed through improving life expectancy of the adult through tackling likely cause of adult death. In Ethiopia adult mortality rate per 1000 population is 487 for males and 422 for females [24].

Though the condition is differing in various studies, adult mortality is high as compared to other countries and still needs special emphasis to reverse the conditions [20]. Among the mortality indicators, cause of death is one of the most highly problematic, especially for developing countries [25]. Tracking change in the basic outcome of adult health is important for assessing progress, improving interventions, and driving further investment [18].

Various studies have been conducted on adult mortality conditions using the BRHP epidemiological surveillance data. Mitike et al. [26], shows that lowland residents encountered the highest mortality rates, and malaria was the most common of the family reported cause of deaths. According to Mitike et al., high mortality among adult was associated with living in the rural areas, particularly in the lowlands and using un-piped water sources for daily life. In addition, the study [26] also states that survival is associated with improving socio-economic conditions and accessibility of the health services in the area.

Another study conducted in Butajira by Yemane et al. [24] reveals that potential markers of epidemiological transition like literacy, source of water, distance from Butajira town, house ownership, age group, sex, and period are risk factors for mortality. The global pattern of adult morality also indicates that that socioeconomic condition is an important determinants of adult survival i.e. the income levels of the populations has direct relationship with adult survival [19].

Accordingly, there are studies using data mining to explore patterns from BRHP data. Shegew [7] applied data mining techniques to predict the risk of child mortality in the area. Amanuel [14] also conducted a research in the area by using data mining techniques in order to predict household health seeking patterns using BRHP dataset. His intention was to develop a model that identifies risk factors and patterns of household health seeking behavior at Butajira district.

Recently, Taddesse [15] also applied data mining techniques to discover knowledge to gain insights in to vital statistics using 18 years BRHP data.

Knowledge from the bulk of data on trends, in age, gender, geographic variations, and burden of disease remains hidden. Inattention of population based information constitutes a major and long-standing constraint on the articulation of effective policies and programs. Therefore, improving the health of the poor through addressing the problem at the source that perpetuates profound inequities in health is becoming vital.

These days, the size of the BRHP database keeps on growing year after year makes it almost impossible to accomplish the desired business objectives using traditional statistical and visualization tools. This is where data mining methods and techniques are becoming useful to health care. One of the greatest triumphs of data mining is anticipating the future by discovering the hidden knowledge with in the huge dataset.

Though varies studies have been conducted using epidemiological tools and data mining technology, all the available knowledge in the area are insufficient to solve the problem of the age-specific adult mortality as it is becoming an important indicator for the comprehensive assessment of the mortality pattern in a population. Therefore, the absence of significant attempts that has been made so far to carry out investigation using data mining techniques on adult age group rationalizes the relevance of this research work.

Therefore, this study aims to construct adult mortality predictive modeling by extracting hidden pattern and knowledge related to adult death based on BRHP open cohort database. Identifying major determinants and risk factors for adult death helps to alleviate adult mortality problem and helps to limit the loss of the productive group.

Hence, this study attempts to explore and answer the following basic research questions.

- What are the major attributes to consider in applying data mining for adult mortality prediction?

- Which data mining technique is more appropriate to construct adult mortality predictive model that can be used in adult mortality prediction?
- What are the optimal variables and determinant factors that lead to adult death in the area of Butajira district?

## 1.3 Objective of the Study

### 1.3.1 General objective

The general objective of this research is to construct adult mortality predictive model using data mining techniques so as to identify and improve adult health status using BRHP open cohort database.

### 1.3.2 Specific objectives

In order to accomplish the general objective, this study has carried out the following specific objectives:

- To understand the problem domain by reviewing literatures and documents on adult mortality, thereby to extract and to prepare the dataset required for mining from the database of BRHP.
- To prepare good quality dataset for analysis by applying preprocessing tasks such as data cleaning, data transformation and attribute selection.
- To build a predictive model using data mining tool and techniques on cleaned BRHP epidemiological data.
- To test the performance of the model using test set and validate mining results with domain experts.

## 1.4 Scope and Limitation of the Study

The scope of the research is delimited to one of the rural health program at Ethiopia centered at Butajira DSS. The research is aimed to apply data mining techniques for discovering significant knowledge using BRHP data and build a model that predicts the status of adult mortality through identifying dominant factors related to cause of adult death in the area.

The inclusion criterion of this study is records of adults whose age group is 15-60 years and the exclusion criterion of this research work is adults whose age group is below 15 years and age

above 60 years. The inclusion criterion follows the agreement of adult age group from 15-60 years by scientific community [4, 17].

The major limitation of this research is, due to time limitation the researcher is unable to apply association rule discovery techniques to investigate the internal association exists among the different variables considered in this research.

## 1.5 Research Methodology

Research methodology explains how the data is collected and analyzed so as to answer the basic research questions. This step of the research contains what research methods are going to be used, the choice of the study design and a strategy of data collection, management and analysis. The method used to attain the desired goal in one particular study is one of the important tools in an academic research.

### 1.5.1 Research design

The study is aimed to uncover hidden patterns in the data records using data from BRHP database. This study follows Hybrid methodology of KDP (Knowledge Discovery Process) [27] to achieve the goal of building predictive model using data mining techniques. Hybrid Process model is selected since it combines best features of CRISP (Cross-Industry Standard Process for Data Mining) and KDD (Knowledge Discovery in Data base) methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives. KDP is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [27].

In addition to these, it is becoming more popular in different settings in knowledge discovery process like medicine and software development areas [27]. For example, development of computerized diagnostic systems for cardiac SPECT images, analysis of data concerning intensive care, cystic fibrosis, and image based classification of cells are some of the area where hybrid model is widely used [27].

As depicted in Figure 1.1 Hybrid methodology basically involve six steps such as problem domain understanding, data understanding, data preparation, data mining, and evaluation and use of discovered knowledge [27].

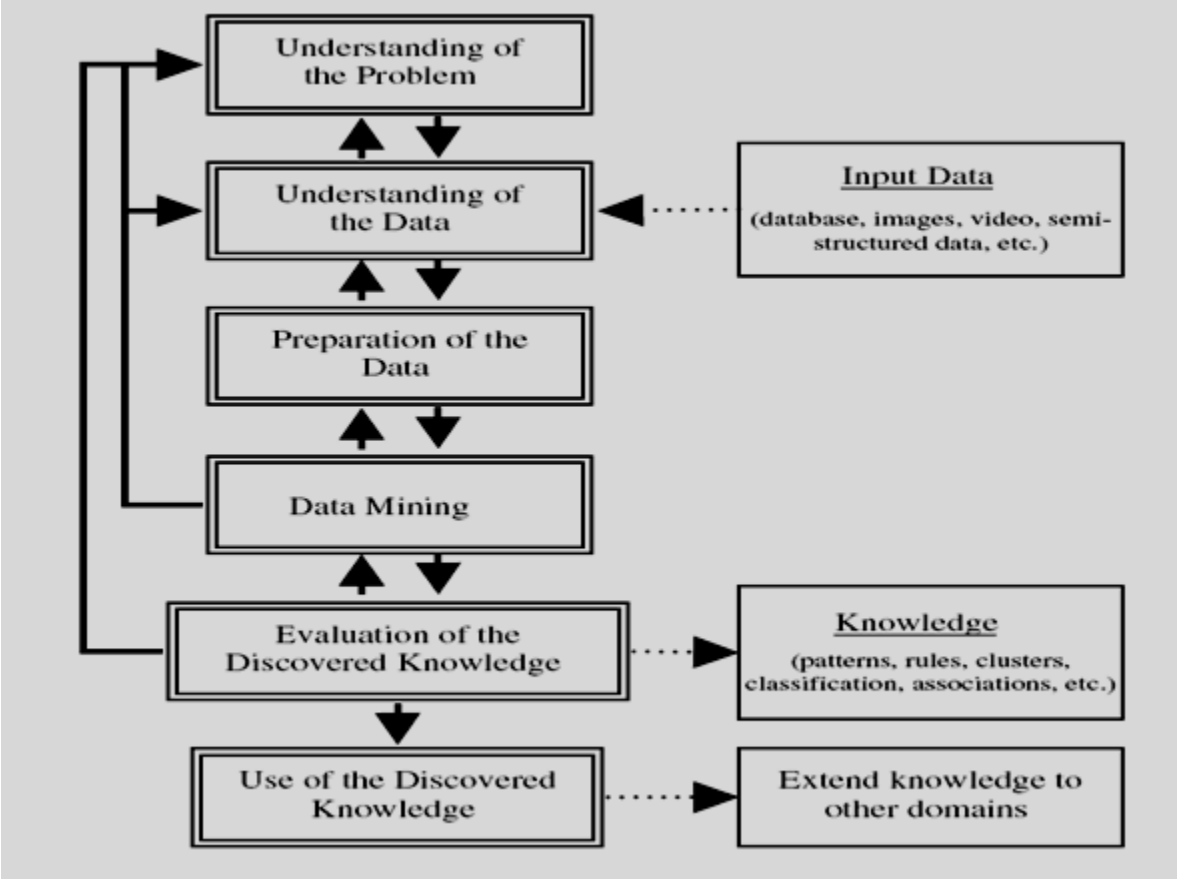


**Figure 1.1 Data Mining Model Used in this Research**

The initial step in the hybrid model is problem domain to define the problem and determine the research goals and learning about current solution to the problem. It also involves learning domain-specific terminology and preparation of a description of the problem, including its restriction. Finally the research goals are translated in to data mining goals and initial selection of data mining tools or data to be used later in the process is performed.

This is followed by data understanding step which includes collecting sample data and deciding which data, including format and size, are needed.

The third step is data preparation which concerns deciding about the data used as input for DM methods in the subsequent steps. The cleaned data may be further processed by feature selection and extraction algorithms. The end result is a good quality data that meets the specific input requirements for the data mining tools selected.

During data mining the data miner uses various data mining techniques such as classification, clustering and association rule discovery to derive hidden knowledge from preprocessed data. This step creates predictive and/or descriptive models. The discovered knowledge is evaluated for understanding the result, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Finally there is a need to plan where and how to use the discovered knowledge. A plan to monitor the implementation of the discovered knowledge is created and the entire research is documented.

Table 1.1 shows the summary of tasks and methods used at each phase of the hybrid methodology.

**Table1.1 Summary of Phases and Tasks of Hybrid Model**

| Phases | Descriptions |
|---|---|
| **Phase 1:** Domain Understanding | Discussion with domain experts has been used as the main source and reviewing different documents, books, journals and articles that focus on data mining techniques in health care were used as supporting sources. |
| **Phase 2:** Data understanding | This stage is creating a target dataset focusing on a subset of variables to get data samples on which discovery aimed to solve the problem. Relevance analysis was done using different des descriptive statistics like data visualization. Data incompleteness, redundancy, missing values and noises in data are observed. Finally, this phase verify the usefulness of the data with respect to the data mining goals. |
| **Phase 3:** Data preparation | This phase is concerned on deciding to make data ready which is used as input for data mining process. To this end, data cleaning (such as filling missing values, detecting outliers) and data transformation are performed using statistical techniques with the help of SPSS tool. |
| **Phase 4:** Data Mining techniques | Data mining algorithms and techniques were experimented for searching interesting patterns and creating predictive model using J48 decision tree algorithm and Naïve Bayes classifier. Weka is used for experimentation. |
| **Phase 5:** Evaluation of the knowledge | Checking whether the discovered knowledge is novel and interesting and interpretation of the results by domain experts. The performance of the predictive model created in this study is also measured using accuracy and ROC (Receiver Operating Characteristic) area. |
| **Phase 6:** Use of discovered knowledge | Different strategies are used to disseminate the discovered knowledge. First: the final report of this research was presented and submitted to School of Information Science and Public Health, Addis Ababa University. Second: the result is also be sent for publication to international and/or local journals for publication. |

## 1.6 Significance of the study

Primarily, the research work has an explicit significance in development of knowledge for the researcher and uses as a benchmark for interested researchers to explore the issues in the area.

Information on adult mortality rates and causes of death is clearly important to inform regional and national health police and other stakeholders who are collaboratively doing on adult health to monitor the impact of interventions and progress towards millennium development goals.

The outcome of the study provides hidden knowledge by extracting large volumes of data and a model uses to predict the risk of adult death by realizing the hidden features from BRHP dataset which contains socio-demographic, causal and other related factors.

Moreover, the study gives a clue on how the death of adults in their prime productive years is affecting household behavior and welfare and also helps to mitigate the impact of adult mortality in the rural community and enhances business goal by indicating where the emphasis of adult health services might be focused to reduce adult mortality by taking proactive knowledge-driven decisions. This further scales up adult health improvement issues through guiding to prevention policies and programmes.

The study has a paramount importance for FMoH (Federal Ministry of Health) and other non-governmental organizations to plan and implement health services focused on adult health strategies so as to mitigate the death attributed by avoidable factors through implementing the extracted rules form the experimentation.

## 1.7 Ethical Considerations

The research does not require personal identifiers like name and ID of the individuals about whom the data is collected.

The research is fully used for academic purpose; ethical clearance was obtained from School of Public Health and Information Sciences. The outcome of the study is supposed to contribute to public health promotion in rural Ethiopia. Hence, the research work does not expose anybody to be harmed in any way.

## 1.8 Organization of the Thesis

The research work is organized in to six chapters. The first chapter deals with aim of health care practice, background of BRHP, introducing the burden of adult mortality conditions, statement of the problem, objective of the study, methodology, scope of the study, limitation of the study and significance of the research work.

The second chapter discusses briefly about data mining and its techniques, methods and algorithms. Application of data mining in health care and some related works with adult mortality condition in relation with burden of adult mortality in developed and developing countries and associated factors with adult death were also addressed in this section.

The third chapter mainly focuses on how the research conducted including what procedures are followed to understand the problem, collect, and analyze the data, build and test the models. Generally, the way how data preprocessed, the model and method used in this research work, and the algorithm selected were discussed in detail in this chapter.

Chapter four attempted the first two steps of hybrid model (problem domain understanding and data understating) to address the driving force of adult mortality through understanding the business area. It also shows the task done to generate the good quality dataset ready to apply data mining tools and techniques. Therefore, preprocessing tasks including data cleaning, transformation and attribute selection are discussed.

Chapter five presents the experimentation done, performance evaluation and the analysis of the result using classification techniques in data mining with selected algorithms

At the end, chapter six provides concluding remarks and recommendation to show further research directions.

# CHAPTER TWO

# LITERATURE REVIEW

Nowadays, data mining application is becoming highly visible tools in many fields like e-business, marketing and retail, medicine and public health. It has led to the popularity of its use in knowledge discovery and extraction of interesting patterns from huge amount of data [28]. Tremendous amounts of medical data are generated every day from individual research efforts, clinical practices, community surveillance and reports from different medical setting like laboratory results and patient records. These data are available in hundreds of public and private databases. Thus, researchers and practitioners are now facing the problem of data overload. These data need to be effectively organized and analyzed in order to extract useful knowledge for sound decision making [28]. New computational techniques are needed to manage these large repositories of data and extract complex relationship in data which is difficult to analyze by traditional statistical tools. This is the main reason why data mining as a new information technology technique is emerged [28].

## 2.1 Overview of Data Mining

Data mining is used intensively and extensively by many fields like agriculture, education, business etc [11]. It has also multi-effects in medicine; for example, it can help healthcare industry to make customer relationship management decisions, physicians to identify effective treatments and best practices, and patients to receive better and more affordable healthcare services [29].

Due to voluminous and too complex amounts of data generated by healthcare transactions in different setting, the application of traditional tools become inefficient to discover useful information from such a huge data. This, in turn calls data mining that enables to explore the information buried in the data and creates models to find hidden patterns in large and complex collections of data which overwhelms traditional methods of data analysis because of the large number of attributes and the complexity of patterns [30].

Though there is some confusion about the term ‘data mining’ with knowledge discovery, and knowledge discovery in databases, many researchers and practitioners use data mining as a synonym for knowledge discovery. Others view data mining as simply an essential tool in the process of knowledge discovery process [27]. There are several definitions for data mining, but the following are the most used ones by the scientific community:

- Data mining refers to extracting or mining knowledge from large amounts of data [11]
- Data miming is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules [11].
- Data mining is the analysis of large observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [10].
- Data mining is the search for new, valuable, and nontrivial information in large volumes of data [12].
- Data mining is a process that uses algorithms to discover predictive patterns which is useful, previously unknown knowledge by analyzing large and complex datasets [31].
- Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models [29].

Many other terms carry a similar meaning to data mining such as knowledge mining from data, knowledge extraction, data or pattern analysis, data archaeology, and data dredging [11]. In summary, data mining is a technology that is used to explore the hidden knowledge/pattern from huge dataset.

## 2.2. Methodology of Data Mining Research

One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets [12]. Data mining tools perform data analysis and uncover important data patterns, contributing greatly to different business strategies including medical researchers. The widening gap between data and information calls for a systematic development of data mining tools that will turn data tombs into golden nuggets of knowledge. Thus, patterns and knowledge from data mining is using for sound judgment and proactive decision making in different organization including health care sectors.

Broadly used methodologies in data mining are KDD (Knowledge Discovery in Data base), CRISP-DM (Cross-Industry Standard Process for Data Mining), SEMMA (Sample Explore Modify Model Assess), and HYBRID process [27, 32].

### 2.2.1 Knowledge Discovery in Database (KDD)

The first KDD process was proposed by Fayyad in 1996 [32]. This process consists of several steps that can be executed iteratively. KDD has been more formally defined as it is non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is the process of knowledge discovery while data mining is a technique applied for knowledge discovery considered as just a step in the entire process [32]. As shown in Figure 2.1, the KDD process consists of five steps: data selection, data preprocessing, data transformation, data mining and interpretation/evaluation.

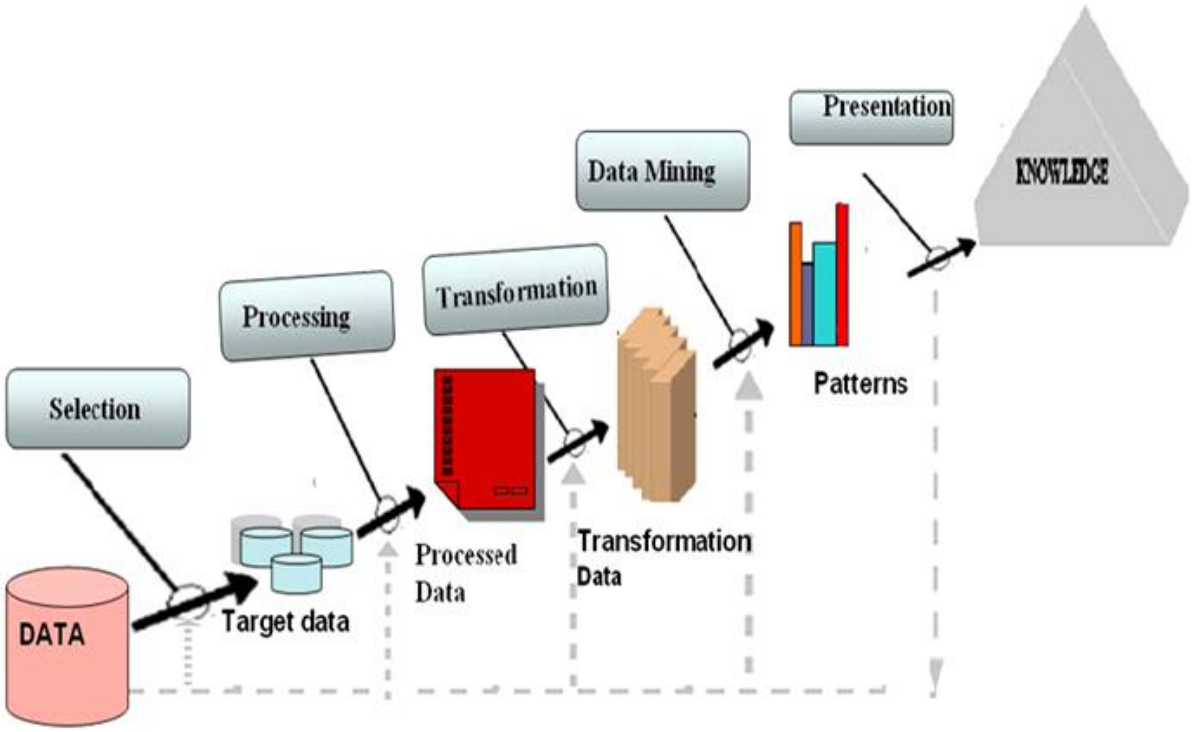


**Figure 2.1 KDD process**

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

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

Because of the use of different sources, data that is fine on its own may become problematic when we want to integrate it. In this step data need to be combined from multiple sources, such as database, data warehouse, files and non-electronic sources into a coherent store. We need to merge different sourced data by keeping uniform format for all before running data mining tools and techniques.

During transformation phase, data are consolidated into forms appropriate for mining to reduce data size by dividing the range of data attribute into intervals each containing approximately same number of samples or to scale attribute data to fall within a specified range. Therefore, values of attributes are changed to a new set of replacement values to ease data mining.

Data mining is the next essential process where intelligent methods are applied in order to extract hidden patterns in the data. This phase requires analysis of the main problem for patterns of interest in the data depending on the business objectives and data mining requirements. Different data mining algorithms and techniques are used for searching knowledge or interesting patterns to construct predictive or descriptive models.

Model creation is followed by performance evaluation which measures the *accuracy* rate of the system. The mined pattern enables to identify the truly interesting ones. For any errors or mismatched result generation as compared to domain area perspectives, the process restarts to initial step so as to provide accurate results.

---

[2] *Accuracy means the percentage of test set samples that are correctly classified by the classifier.*

Finally, visualization and knowledge representation are used to present the mined knowledge to the users and stored as new knowledge in the knowledge base. Incorporating the knowledge in to another system for implementation purpose, documentation and report for presenting the benefit of the knowledge to interested parties, incorporating the knowledge with previously known knowledge in the area are some of the important activities during this phase.

### 2.2.2 CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP–DM was developed in 1996 by analysts for fitting data mining into the general problem solving strategy of a business or research unit [29]. CRISP–DM is one of the most widely used methodologies in extraction of knowledge which has a life cycle consisting of six phases which is an iterative and adaptive process [27], as depicted in Figure2.2

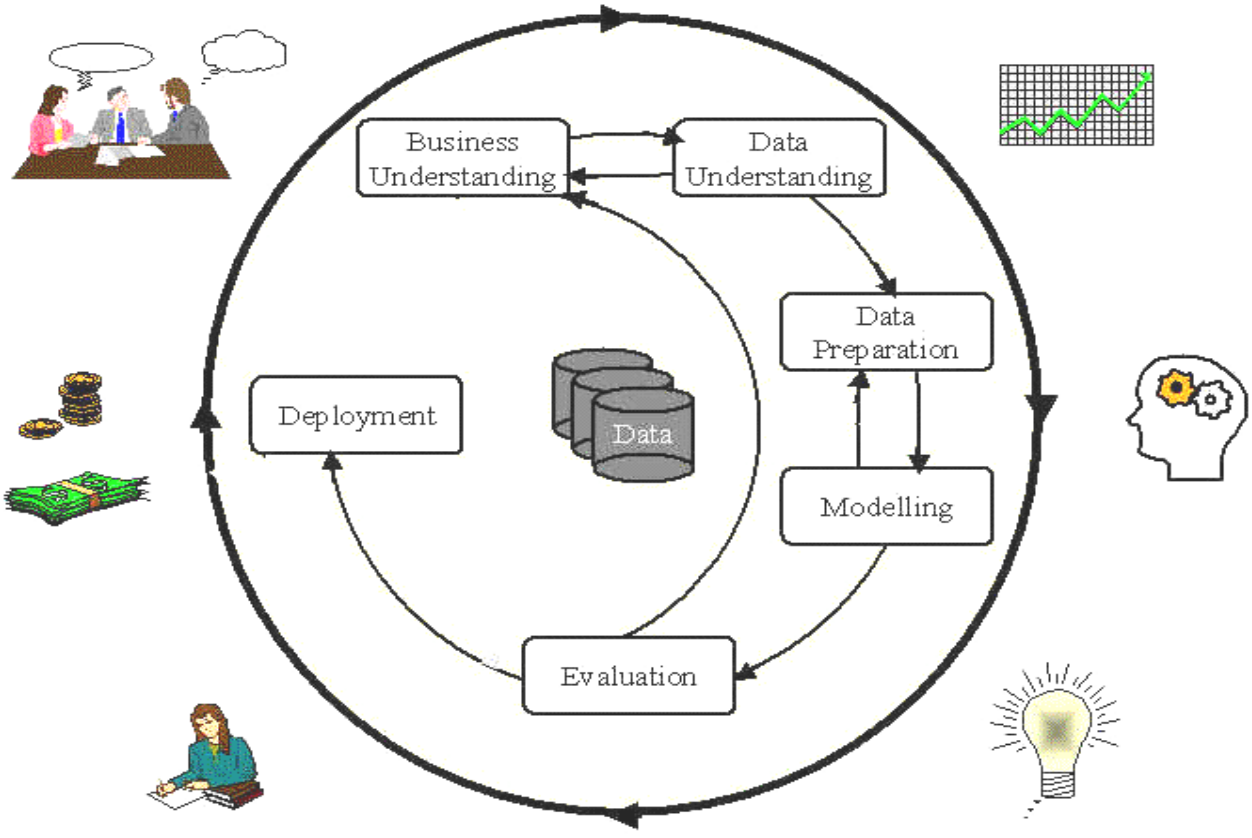


**Figure 2.2 The CRISP-DM knowledge discovery Process Model**

In CRISP- DM, the sequences of the phases are adaptive i.e. the next phase in the sequence often depends on the outcomes associated with the preceding phase.

Business understanding is the initial phase in the CRISP–DM standard process which focuses on understanding business area in which DM objectives and project requirements are assessed as a whole from business perspective points of view.

It also translates these goals and objectives into the formulation of a data mining problem definition and prepares a preliminary strategy for achieving the desired objectives. Further, it is broken in to determination of what clients really wants to accomplish from business perspectives, assessment of the situation for fact finding about the resources, constraints and assumption, determination of data mining goals, and states project objectives in technical term and finally description of the project plan for achieving the data mining and business goals.

Once the business is well defined and understood, data understanding phase begins with collecting the initial data and continues with several activities in order to become familiar with the data that helps to identify quality of the data. During this step, the following tasks are performed. First, collecting initial data for modeling; second, data description to get insights through descriptive statistics available in the statistical tools; third, exploration of data to capture an overall sense of the dataset through computing summary. Lastly, verification, and visualization of data quality is checked if any unnecessary data fields with incomplete, inconsistent, noisy, and redundant values existed.

The data preparation step contains all activities needed to construct the final dataset. It starts from preparing the initial raw data to the final dataset which is ready for application of data mining tools. This step further divided in to four steps. The first step is data selection that is appropriate for analysis. This is followed by data cleaning which is making data ready for the modeling tools. Data construction is the third step that attempts to produce derived attribute, new records and transformed values for existing attributes. Then, data integration is to combine data from multiple sources (records or tables), and, finally data formatting for reconstructing data values without changing its meaning.

After data being ready to apply data mining tools in proceeding step, various modeling techniques are selected and applied at this phase. Since some data mining tools may require specific formatting for input, it may needs reiteration into the previous phases for improvement.

The modeling step selects first modeling techniques based on data mining objectives and also generates test set to evaluate and validate model performance. This is followed by model building using the modeling tool. Finally, assess and interpret the pattern according to the domain knowledge.

After models have been built, they need to be evaluated to check whether they fulfill the requirements and objectives set at the beginning of the project. The model is also evaluated from the point of business objectives. Reviewing of steps executed to build the model and evaluating the models for quality and effectiveness before generating them for end users in the field are also performed. At the end, decision regarding the deployment and use of the data mining results is reached.

### 2.2.3 SEMMA

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

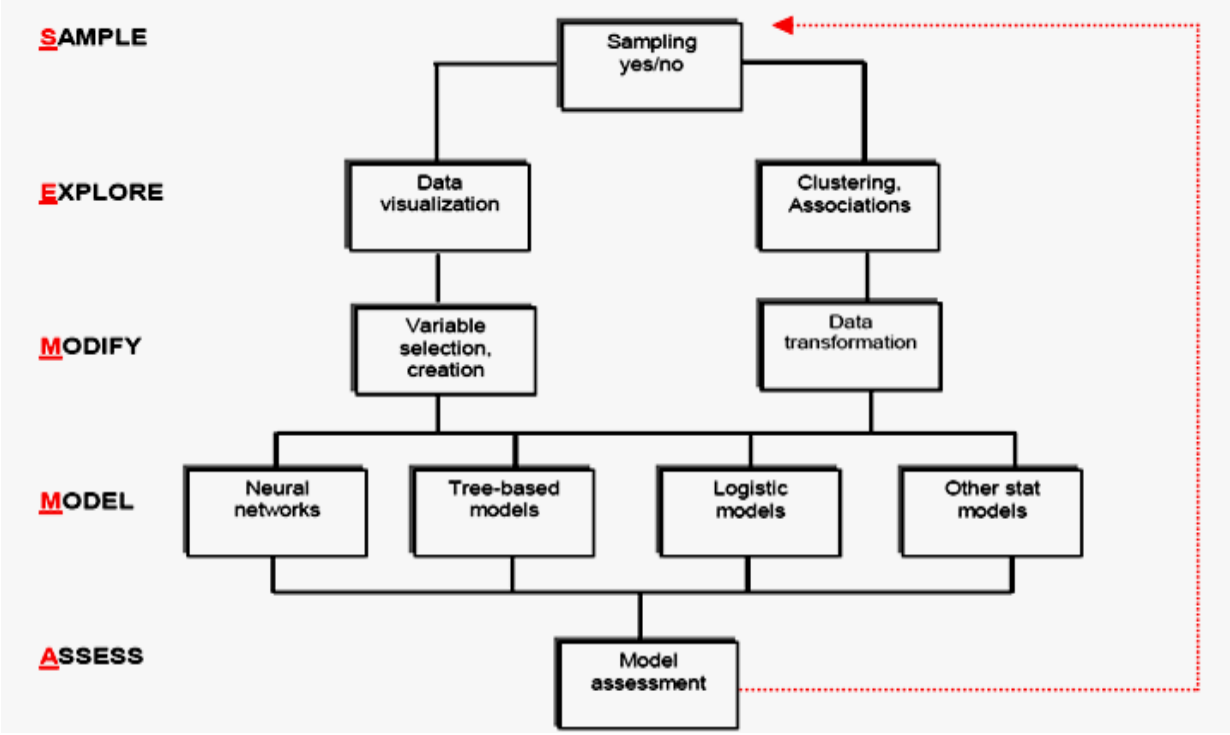


**Figure 2.3 SEMMA Process model**

The first phase in SEMMA process model is sampling. In this phase, a portion of a large dataset is extracted in order to take reliable and statistically representative sample from the huge data for optimal cost and computational performance.

Sample data selection is followed by explore. This is a state where searching for unanticipated trends and anomalies in order to gain a better understanding of the dataset occurs. This helps to refine the dataset and redirect the discovery process.

During the modify phase, user creates, selects, and transforms the variables upon which to focus the model construction process. In addition, it manipulates data to include information and handle outliers to increase significance of variables and focus on model selection process.

Once the data is prepared, the modeling phase constructs models that explain patterns in the data by applying modeling techniques in data mining. The modeling techniques are selected based on the objectives of the data mining project.

This is the assess phase in which the usefulness and reliability of the model is evaluated from the data mining process and how well it performs. A common means of assessing a model is to apply it to a portion of dataset put aside for testing during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, one can test the model against known data.

By assessing the outcome of each stage in the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. In SEMMA, the sample steps goes equivalently with selection step of KDD and continues till to last assessment phase as interpretation/evaluation of the discovered knowledge in KDD. However, KDD manifests the pre-KDD and Post KDD that SEMMA does not.

### 2.2.4 Hybrid model

The development of academic models such as the nine-step model and eight-step model and industrial models such as five-step model and the six-step CRISP-DM model has led to the development of hybrid model that combines aspects usable for DM research. It was developed by Cios et al. [33] based on the CRISP-DM model.

Hybrid process is characterized by providing more general, research oriented description of the steps. The hybrid model also encourages the application of knowledge discovered for a particular domain in other domains and it has a six step process as depicted in Figure 1.1 [27].

Summary of correspondences between KDD, SEMMA, CRISP-DM, AND HYBRID models are presented in Table 2.1 [32].

**Table 2.1 Summary of data mining models**

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

From the Table 2.1 by doing a comparison of the models, some of them follow same steps to discovery process while others follow different steps. For example in KDD and SEMMA stages the first approach is equivalent. Sample can be identified with Selection; Explore can be

identified with Pre processing; Modify can be identified with Transformation; Model can be identified with Data Mining; Assess can be identified with Interpretation/Evaluation.

## 2.3 Data Mining Function

Data mining is utilized for the intention of finding of hidden information in a database upon developing of model which could best fit the data. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. The ability to extract useful knowledge hidden in the data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer based methodology including new techniques for discovering knowledge from data is core function of data mining. It searches for new, valuable, and nontrivial information in large volumes of data [12]. Data mining tasks are in general classified in to two main categories [12]: predictive-oriented and descriptive oriented.

Predictive data mining tasks produce the model of the system described by the given dataset to build a model that permits the value of unknown variable to be predicted from the known values of other variables [10]. It is a technique that involves using some variables or fields in the dataset to predict unknown or previously unseen future values of other variables of interest. It is usually used to create a model based on a set of predictors to relate the dependent variables. Examples of predictive modeling includes classification, prediction etc.

The second category of data mining function is descriptive mining task. This is another data mining task used to characterize the general properties of the data in the database [11. It produces new, nontrivial information based on the available dataset and is to gain an understanding of the analyzed system by uncovering patterns and relationships in large datasets. The goal of a descriptive model is to describe all of the data or the process generating the data [10, 12]. Examples for descriptive data mining are clustering, summarization, association rule discovery, and sequence discovery. The followings are some of the examples from both data mining tasks how they are working in real pattern discovery process.

### 2.3.1 Classification

Classification is one of the predictive data mining tasks. It is a technique used to predict group membership for data instances by assigning previously unseen records a class as accurately as possible. It is said to be the process of finding a model or function that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown [11].

The derived model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision trees, mathematical formulae, semantic network etc [11]. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attributes set and the class level of the input data.

After having an accepted accuracy level, one can use the model for classification of new data tuples. Applications examples of classification in health sectors are the following [11].

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

There are various classification algorithms; among which the main ones are the following [11].

#### 2.3.1.1 Decision tree induction

When decision tree induction is used for attribute subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples. Thus, the algorithm choose the best attribute to partition the data into individual classes includes ID3, C4.5, and CART [12].

In decision tree construction, selection of splitting attributes is necessary in order to avoid irrelevant attributes by examining the effect of each attribute for the distinct class and its likelihood for improving the overall decision performance of the tree, since the feature with minimum impact on dependent variable may distort the trees performance and the classification accuracy.

There should be certain requirements before decision tree algorithms become applied [28]. First: since decision tree algorithms represent supervised learning, they require pre-defined target variables and training dataset which provides the algorithm with the values of the target variable. Second: this training dataset should be rich and varied, providing the algorithm with a healthy cross-section of the types of records for which classification may be needed in the future.

Decision trees learn by example, and if examples are systematically lacking for a definable subset of records, classification and prediction for this subset will be problematic or impossible. Third: the target attribute classes must be discrete i.e. one cannot apply decision tree analysis to a continuous target variable. The target variable needs to take on values that are clearly demarcated as either belonging or not belonging to a particular class.

One of the most attractive aspects of decision trees lies in their interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf [28]. Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule [12].

The challenge with decision tree is overfitting. As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex [11]. This potentially leads to the concept of overfitting which consequently brings the notion of pruning; this implies removing of branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree [11].

#### 2.3.1.2 Rule based classification

Though the decision tree is a widely used technique for classification purposes, another popular alternative to decision trees is classification rules which can be expressed as paths IF-THEN rules so that humans can understand them easily [34]. A rule-based classifier uses a set of IF-THEN rules for classification; it is a relationship between antecedent, and consequent i.e. an expression of the form IF condition THEN the conclusion.

The algorithm decision tree is the best known method for deriving rules from classification trees [35]. For example, one could have the following set of rules to classify the weather condition. If temperature < 50°F, then weather = cold. If temperature > 50°F AND temperature < 80°F, then weather = warm. If temperature > 80°F, then weather = hot [36]. Although any of the logical expressions are allowed, preconditions are usually connected with the AND operation.

The advantage of IF-THEN rule is the rules are order independent i.e. regardless of the order of rules executed, the same classification of the classes is possible to reach [36]. The challenges is the generated rules are often more complex than necessary and contain redundant information and the rules generated this way may be unnecessarily complex and incomprehensible [36].

#### 2.3.1.3 Neural network

It is represented as a layered set of interconnected processors which has a relationship with the neurons of the brain with weighted connections between the units. An individual node take the input received from connected nodes and use the weights together to compute output values and learns by adjusting the weights so as to be able to predict the correct class label of the input tuples [11].

The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of neuron like units, which is hidden layer.

The outputs of the hidden layer units can be input to another hidden layer depending on the NN [3]*architecture* [37]. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples as depicted in Figure2.5 [11, 35]. Neural networks are quite robust with respect to noisy data and generating a nonlinear response as well as their ability to classify patterns on which they have not been trained. Neural networks have been criticized that they involve long training times and for their poor interpretability since it is difficult for humans to interpret the symbolic meaning behind the learned weights; these features initially made neural networks less desirable for data mining. Therefore, it becomes more suitable for applications where this is feasible [11].

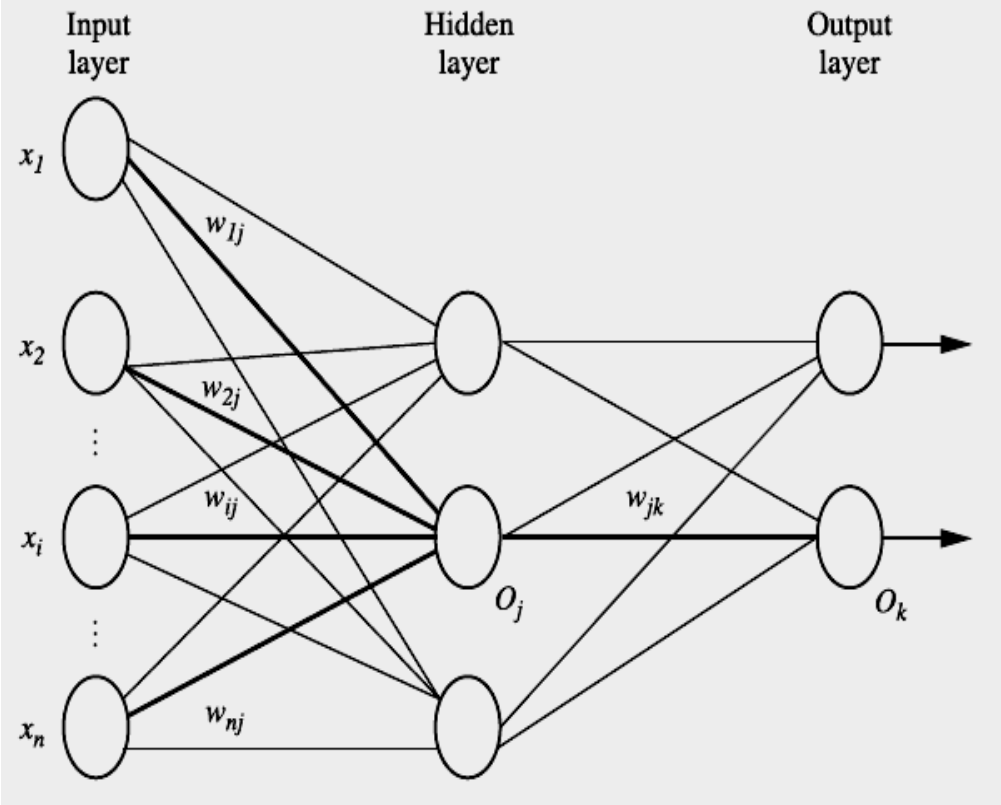


**Figure 2.5 Simple Feed Forward Neural Network**

[3] *The architecture of the neural network is the specific arrangement and connection of the neurons that make up the network.*

#### 2.3.1.4 Naïve Bayes classifier

Bayesian classifier is statistical classifier and a practical learning algorithm that can predict class membership probabilities. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes and classification is based on a probabilistic model specification; i.e. it can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class [11].

When using the Naïves Bayes method to classify a series of unseen instances the most efficient way to start is by calculating all the prior probabilities and also all the conditional probabilities involving one attribute though not all of them may be required for classifying any particular instance. Naïve Bayesian classifier assumes class conditional independence i.e. it treats each variable independently and measure the effect that different values of the variables [38].

The advantage of Naïve Bayes classifier is it reaches the minimum error when the dataset is large and the methods for estimating particular probabilities are consistent [36]. However, the challenge in Naïve Bayes classifier is that the model used in the classification might not be the best estimator of the probability distribution though it has multi-effects in different area where it has relatively good performance [36].

In Bayesian network (Belief network) which is the graphical model of causal relationships that represent dependency among the variables and it gives a specification of joint probability distribution so that it is used to solve the variables interdependences [39].

### 2.3.2 Clustering

Clustering is concerned with grouping together objects that are similar to each other and dissimilar objects belonging to other clusters [35]. Clustering data mining algorithms is useful for exploring data, and used to find natural groupings [30]. In many fields, there are obvious benefits of grouping similar objects together, like in an economics, financial application, marketing and crime analysis application. Also in medical area clustering are used to group patients according to their similar symptoms [35].

In machine learning, clustering is considered as an unsupervised learning. Unlike classification, clustering does not rely on predefined classes and class-labeled training. It focuses on observation rather than learning by examples [11]. Clustering algorithms seek to segment the entire dataset into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity of records outside the cluster is minimized [35].

There are two most commonly used clustering approaches: k-means clustering and hierarchical clustering [35]. K-means clustering is an exclusive partitioning clustering in which each object is assigned to precisely one of a set of clusters that the user would like to form from the data given [35]. The K-means partitioning clustering algorithm is the simplest and most commonly used algorithm employing a square-error criterion. A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function if the initial partition is not properly chosen [12].

In hierarchical clustering, a treelike cluster structure is created through recursive partitioning or combining of existing clusters [12]. The quality of a set of clusters is determined using the value of sum of the squares of the distances of each point from the centroid of the cluster to which it is assigned [35].

The quality of a cluster is represented by its diameter which is the maximum distance between any two objects in the cluster [11].

### 2.3.3 Association rule discovery

Association rules are one of the major techniques of data mining in unsupervised learning system. While classification and clustering are global pattern discovery of DM task, association rule discovery is the most common form of local-pattern discovery [12]. Of the data mining tasks, association rule mining method is applied in either supervised or unsupervised manner [28]. Association rule mining searches for interesting relationships among items in a given dataset and patterns are represented in the form of association rules [11]. Association rules such

as the occurrence of some items in a transaction will imply occurrence of other items in the same transactions [40].

Interestingness of patterns need to be measured to make sense that patterns are easily understood by humans, valid, potentially useful, and novel [11]. In the case of classification rules, we are generally interested in the quality of a rule set as a whole. It is all the rules working in combination that determine the effectiveness of a classifier, not any individual rule but, in the case of association rule mining the emphasis is on the quality of each individual rule [24].

There are several objective measures of pattern interestingness exist, like support and confidence. Support is the percentage of transactions from a transaction database that the given rule satisfies and confidence assesses the degree of certainty of the detected association. For example let us see how the two measures of pattern interestingness work [11].

Support (X➔Y) ==>P(X U Y); meaning the probability that a transaction containing *X* also contains *Y*.

Confidence (X➔Y) ==>P(Y /X); meaning the probability that a transaction containing *X* also contains *Y*.

### 2.3.1 Apriori Approach to pattern mining

Association rule mining is also one of the pillars of data mining used to uncover relationships between inherently unrelated data items. It is an implication that one item is associated with another item or one disease occurrence is associated with another disease occurrence.

Formally, given a set of m items I = {$I_1$, $I_2$,…$I_M$} and a database of n transactions D = {$t_1$, $t_2$,…, tn}, where a given transaction contains k items $t_i$= {$t_{i1}$,$t_{i2}$,… ,$t_{ik}$} and $I_i$ ϵ I, an association rule is an implication of the form X➔Y, where X,Y ϵ I are sets of items and $X \cap Y = \emptyset$ [36].

In data mining there are different itemset-mining algorithms such as Apriori [36].

The Apriori algorithm computes the frequent item sets through several iterations; each iteration having two steps [12]. These are candidate generation and candidate counting and selection that fit the minimum support requirement to next pass. The algorithm uses prior knowledge of

frequent item set properties and states that, non-empty subsets of a frequent item set must also be frequent [36,38].

In medical health care, there are certain diseases occurrences that commonly associate with other diseases. For example, in most of the cases in health care tuberculosis follows HIV/AIDS, hypertension follows heart failure, diarrheal disease followed malnutrition visa versa etc. Therefore, association rule mining helps to prove such usual happening conditions whether it is real from the business perspective and to discover other events that occur in hidden fashion [41].

From an application perspective, association rule discovery will be a base for taking a measure depending on the rules with high confidence and high support to ease the sale transaction [36]. The challenges in Apriori is that it assumes that the entire database is memory resident, which may be problematic for situations involving large amounts of data and the maximum number of database scans is one more than the size of the largest itemset, resulting in a large number of scans and slower the performance [36].

### 2.3.2 FP-Growth approach to pattern mining

Frequent pattern growth method is an efficient way of mining frequent item sets in large databases [12]. The algorithm mines and finds frequent item sets without the time consuming candidate generation process. FP-growth first performs a database projection of frequent items then mining the main memory by constructing a compact data structure which is FP-tree [12].

Start from each frequent length-1 pattern and construct its conditional pattern base which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern, then construct its conditional FP-tree, and perform mining recursively on such a tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree [11].

A study on the performance of the FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm [11].

## 2.4 Related Works

Adult mortality remains poorly measured in many developing countries since registration of death is incomplete and information regarding age is often inaccurate [5]. As cited in Rajaratnam et al.[33], Feachem and colleagues drew attention to deaths in adults aged 15–59 years and stated that death in the most economically and socially active groups can also have major effects on society. A study conducted based on a database of 3889 measurements of adult mortality for 187 countries from 1970 to 2010 using vital registration, census and survey data for deaths in the household states that, adult mortality varied substantially across countries and overtime [33].

According to Feachem and colleagues [33], in Iceland, the probability of male death before $60^{th}$ birth day is 65 per 1000; in Cyprus, the probability of female death before $60^{th}$ birth day is 38 per 1000. The analysis shows that in 1970, there are 430 women adult death per 1000 and 583 male death per 1000 before reaching their $60^{th}$ birth day. However, in 1990, death between male and female are inclined to 440 deaths per 1000 for women and 617 deaths per 1000 for male before their $60^{th}$ birth day. In 2010, 372 female death per 1000 and 484 male death per 1000 before $60^{th}$ birth day was occurred [33]. In all the cases**,** the effect has been larger on male mortality than it has on female mortality (480/1000 and 372/1000) respectively.

According to Kitange HM [3], a study in rural Tanzania showed that age specific mortality for adults was as much as 43 times higher than rates in England and Wales. In some African countries example Sierra Leone, the adult mortality risk is more than 50% [3]. Hill states that [3] adult mortality in Mongolia, over 50% of the females who survive to 15 years die before 60 years whereas the corresponding risk for females in the Republic of Korea is only 7%. Study in Senegal indicates that an estimated probability of dying between ages 15 to 60 years is around 51.8% for males and 43.7% for females [25].

Adult mortality from communicable and reproductive diseases appears to be much higher in sub-Saharan Africa than elsewhere in the world [25]. The three broad groups of causes of death were communicable and reproductive diseases, non-communicable diseases and injuries [25].WHO estimates of causes of adult mortality [42] from 15 to 59 years shows that in sub- Saharan Africa communicable diseases and maternal deaths are much higher than non-communicable diseases,

whereas in Southeast Asia deaths due to non-communicable diseases are higher than communicable diseases.

A study in Vietnam and Korea reveals that educational attainment is one of the important factors for survival and it was related to mortality in most causes of death [3]. Another study in rural Italy reported that men with college education were found to have significantly higher survival rates as compared to men who have no formal education [3]. According to a demographic model that has been developed for South Africa [43], over 70% of the death among the 15 to 49 year olds can be attributed to AIDS.

In 2006 FMoH AIDS report [44] estimated that AIDS accounts for 34% of all the deaths of adult age form 15-49 years in Ethiopia. Mitike et al. [26] shows that the most important socio-demographic factors that significantly associated with adult mortality are having no educated person in the family OR 1.91; 95% CI 1.11, 3.29), the male sex (OR 1.46; 95% CI 1.09, 1.95), and living in the rural lowlands (OR1.54; 95% CI 1.03, 2.31).

A study that was conducted using a multivariate regression model suggested that young adults from the rural highlands and lowlands had a higher risk of death (adjusted rate ratios 1.99 [1.40-2.83] and 2.58 [1.82-3.66] respectively than young urban adults [45].

By using the information gathered by the BRHP epidemiological surveillance system, several studies were conducted. Yemane et al. [5] shows an interesting pattern where the excess male to female mortality in the urban area for all age groups while, in the lowlands, there is a shift to the disadvantage of females above 45 years of age. The rural to urban mortality differentials are somewhat more pronounced for females, especially for adults. And stated that the leading perceived causes of death were: malaria, diarrhea, tuberculosis, other causes and unknown causes. According to Yemane et al., attributes like period, source of water, literacy, type of house, sex, residential area and age are factors that cause disease entity in the area.

One study [46] shows that the major causes of death were acute febrile illnesses (25.2%), liver diseases (11.3%), diarrheal diseases (11.1%), tuberculosis (9.7%) and HIV/AIDS (7.4%). Overall communicable diseases accounted for 60.8% of the deaths. The high levels of mortality from communicable diseases reflect the poor socioeconomic development of the country, and the general poor coverage of health and education services in rural Ethiopia.

According to Yemane et al. [24] potential markers of epidemiological transition like literacy, source of water, distance from Butajira town, house ownership, together with age group, sex, and period were examined as risk factors for mortality which allow for individual person time contributed in each residence episode.

In the area not only statistical methods were used to investigate the novel and interesting patterns, but data mining technologies also were applied to discover hidden knowledge that might be helpful for business decision making purpose.

Shegaw [7] applied data mining techniques to investigate the potential applicability of data mining technology to predict the risk of child mortality based up on community based epidemiological dataset gathered by the BRHP epidemiological study. Shegaw used a sample dataset consisting 1,100 records taken randomly from the two classes of children (i.e. alive and died) of the ten years surveillance dataset of the BRHP epidemiological study which contains a total of 64,077 records. To build predictive models he used neural network and decision tree techniques and the performances were 93% and 95% respectively. He stated that decision tree approach provided simple rules that can be used by nontechnical health care professionals to identify cases for which the rule is applicable.

The researcher has also identified public health and socio demographic determinants that are associated with infant and child mortality in rural communities [7].

Another study has been conducted by Amanuel [14] using data mining techniques to predict household health seeking patterns using BRHP dataset. The researcher aim was to develop a model that identifies risk factors and patterns of household health seeking behavior at Butajira district. He used a total of 60,446 records for experiments with implementation of J48 decision tree techniques. The finding of the researcher indicated that with an accurate rate of 89.9017%, predicting household health seeking pattern through data mining techniques is possible.

Taddesse [15] applied data mining techniques to discover knowledge that can be used to gain insights in to vital statistic aspects based up on community based epidemiological dataset gathered by the BRHP Epidemiological study. The researcher extracted 18 years data which

contains a total of 236,549 records of which he used 95,220 cases to build predictive model using classification algorithm such as J48 to extract interesting knowledge from temporal data on BRHP data base. He used J48 algorithm that over 90% accurate results are possible for developing classification rule that can be used in prediction. From this result the researcher concluded that prediction is possible using vital statistics data through the application of data mining classification techniques.

In all the researches done, scholars tried to search a new knowledge that may help for business objectives in relation with magnitude and risk factors for adult mortality using BRHP data as an information source on the base of different epidemiological methods like SPSS, EPI-Info, and STATA etc. Using such tools become inefficient to detect unanticipated interesting patterns from voluminous data. In addition, most of the studies have been stayed at a quantitative feature studies with vital information, without considering various bias effects of the data. For utilization of relevant information which is hidden in the data, it is obvious that one need to be engaged in information computational management (data mining technology) since it is efficient to find unrecognized new knowledge and can mine the knowledge rules automatically from the content of data. Therefore, the researcher enthused to prepare predictive model based on BRHP that predict adult mortality pattern.

Previously there are researches works that have been carried in application of data mining techniques using BRHP data as an important information source. But, to the knowledge of the researcher, no previous researches have been done to predict adult mortality by applying data mining techniques in the area.

In scaling-up the proven effective interventions for the target of adult health promotion and prevention through improved health care service by applying data mining technology is very vital. Thus, this research has a great contribution to generate patterns that help in planning a better strategy and effective decision making for adult health promotion plans and programs.

# CHAPTER THREE

## TECHINIQUES FOR MINING BRHP DATA

The overall research design is to build a model that predicts the status of the adult that is the probability of he/she alive or die. Hybrid methodology was also followed to explore the application of data mining on Butajira rural health program. WEKA 3.6 data mining tools and techniques are utilized as means to address the research problem.

Though data mining is newly emerging discipline to handle enormous data which is usually difficult to analyze through traditional tools and techniques, it has different methods and techniques that makes the data mining more popular in different settings including health care.

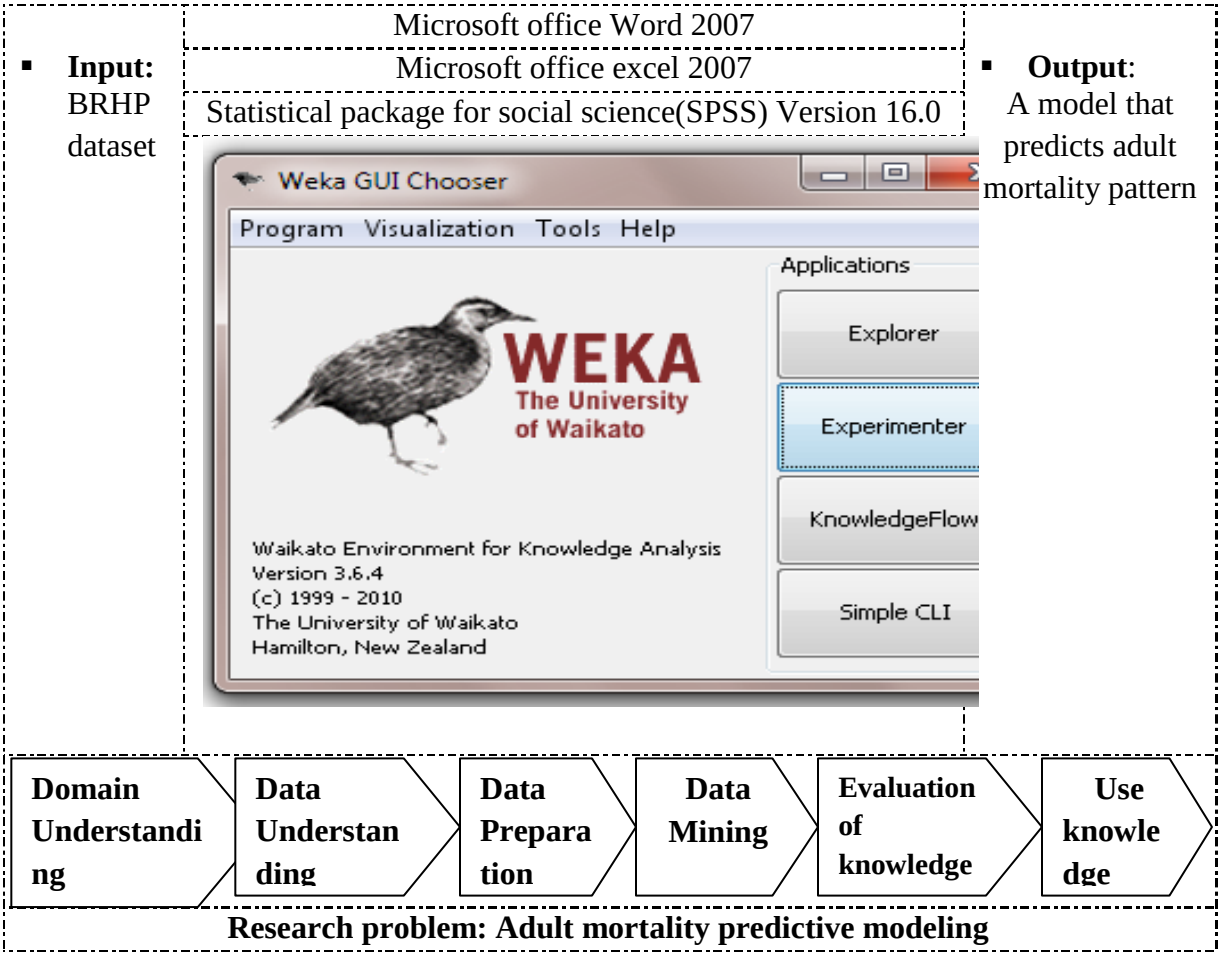


**Figure 3.1 Diagrammatic Overview of the Overall Research Design and Methodology**

## 3.1 Methods of Problem Domain Understanding

This initial step has been thoroughly attempted to understand the driving force of adult mortality that need to be addressed. To accomplish this target, various tasks have been performed such as closely working with domain experts in order to define the problem and determine the research goals, identifying key people and learning about current solution to the problem, learning domain-specific terminology and preparation of a description of the problem are considered as a means of solving the problem.

## 3.2 Methods of Data Understanding and Data Preparation

The task precede DM step such as data cleaning and data transformation were performed in order to get the consistent data. It concerned on deciding to make data ready which is used as input for data mining process in subsequent steps. To develop first insight into the data, relevance analysis like descriptive data visualization was done using statistical tool (SPSS16.0)

In this particular research, adult ages 15-60 years were considered from the entire database. Using stratified random sampling technique, 43,864 study subjects were selected from the ten peasant associations of the district.

As stated by Han and Kamber [11], classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. In order to handle different data related problems in BRHP dataset like missing values, noises, and irrelevant features were solved using tools like Microsoft Excel, SPSS, and Weka 3.6. Exporting data from existing format (SPSS 16.00) to Weka understandable format like arff and csv were also employed.

The data in the existing classes are not equally distributed i.e. there are imbalance data between the classes died and alive. According to Han and Kamber [11], if one class of the target attribute has much lower relative frequency than the other class, balancing this class is recommended. Unless data in classes being balanced, the classification model could simply predict for some class that have more relative frequency to all operation. Therefore, SMOTE technique following the explorer preprocess menu bar was implemented in Weka so as to balance the target attribute (died, alive).

## 3.3 Methods of Modeling

These days, data mining is more popular in different settings including health care areas by mining knowledge from the massive repositories. Commonly used techniques for data classification and prediction in data mining are decision tree induction, Bayesian classification, rule-based induction, the neural network support vector machines, and *k*-nearest neighbor classifiers [11]. However, for this particular research problem, classification algorithms such as decision tree induction (J48) and Naïve Bayes classifier are selected.

### 3.3.1 J48 Decision tree algorithm

Decision tree algorithms have predictive performance ability and capability to discover patterns in huge datasets and understandability of the generated rules by human. For example, the acquired knowledge in tree form using decision tree takes less mental strain to understand the path from the root to leaf and one can generate rule from the tree in order to predict the class for unknown records. In addition rule assimilation easily by end users, classification steps of decision tree induction is simple and fast, and also tree construction does not require any domain knowledge [11].

Therefore, Weka software based decision tree algorithm (J48) was used which is a greedy algorithm i.e. it constructs trees in a top-down recursive or divide-and-conquer manner and orders the class rule sets so as to minimize the number of false-positive error [28]. It uses the concept of information gain or entropy reduction to select the attribute with the highest information gain.

Suppose that we have a variable X whose K possible values have probabilities $P_1$, $P_2$….$P_k$, the smallest number of bits, on average per symbol, needed to transmit a stream of symbols representing the values of *X* observed is the entropy of X. Entropy is the expected information needed to classify a tuple in X:

$$H(X) = -\sum Pj \log 2(pj) \qquad (3.1)$$

For an event with probability *p*, the average amount of information in bits required to transmit the result is $-\log_2 p$. For variables with several outcomes, we simply use a weighted sum of the $\log_2 pj$'s, with weights equal to the outcome probabilities. Therefore, the mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows [28].

$$Info_A(X) = \sum_{j=1}^{v} \frac{|X_j|}{|X|} \times Info(X_j) \tag{3.2}$$

Information gained by branching on attribute A is

$$Gain(A) = Info(X) - Info_A(X) \tag{3.3}$$

At each decision node, C4.5 uses the attribute with the maximum gain ratio as the splitting attribute and recursively visits each decision node, selecting the optimal split, until no further splits are possible. J48 also used the same concept to construct the decision tree and it supports both numeric and nominal predictors and nominal class attribute. It has the capability to handle missing values in datasets [11]. Once the tree is constructed, it is possible to generate the rule in order to apply it for new instances which are independent of the training tuples. Figure 3.2 illustrates the decision tree constructed from which one can easily generate the decision rule.

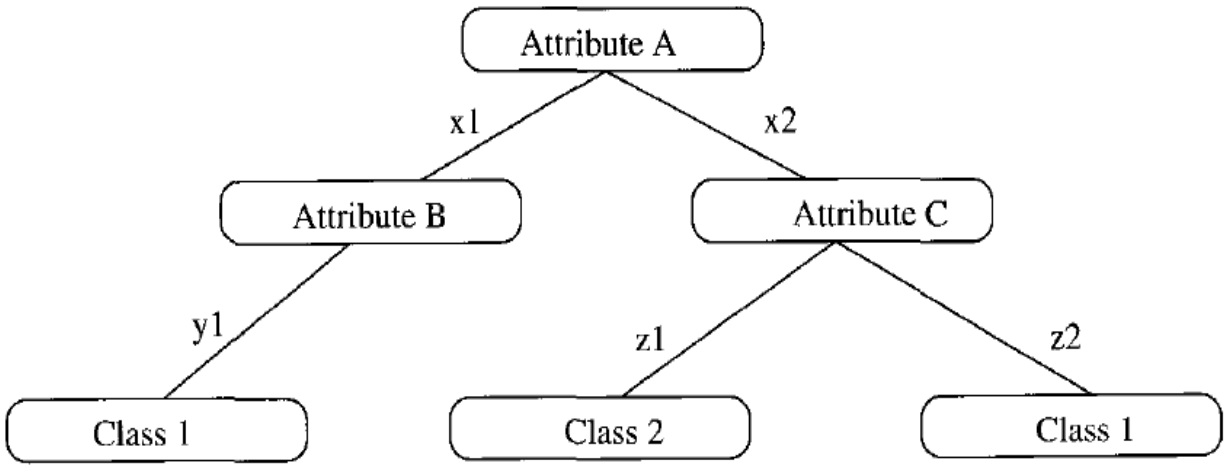


**Figure 3.2 Simple Decision Tree Constructed for Two Class Classification**

From the above simple decision tree, the following rules can easily be generated as follows.

- Rule 1: If (A = *X1* and B = Y1), then Classification = Class 1;

- Rule 2: If (A = X2 and C = *Z1),* then Classification = Class 2;
- Rule 3: If (A = *X2*and C = *Z2),* then Classification = Class 1.

Conditions for stopping partitioning includes all samples for a given node belongs to the same class and there are no remaining attributes for further partitioning.
However, when decision trees are built, many of the branches may reflect noise or outliers in the training data. Considering the goal of research and improving classification accuracy of unseen data, tree pruning was attempted in order to identify and remove unnecessary branches that lead to over-fitting of the model.

The second algorithm is Naïve Bayes classifier which has found to be comparable in performance with other algorithms in data mining like decision tree and neural network classifiers [11]. Naïve Bayesian classifier has also exhibited high accuracy and speed when applied to large databases and it also assumes that the effect of an attribute value on a given class is independent of the values of the other attributes [36].

### 3.3.2 Naïve Bayes classifier

Naïve Bayes classifier is one of the algorithms that produces the lowest classification error rate on a validation data and produces high accuracy performance as compared with others algorithms in classification and prediction in data mining [10].

In general, Naïve Bayes follows the following steps for classification purpose: first, Collect data and estimate parameters such as mean and covariance for each class. Second, choose a set of features that a classifier uses to compute a posteriori probability. Third, choose a model to derive a decision rule with these parameters. Fourth, train the classifier to test dataset and classify each sample. Finally evaluate the decision rule in order to improve the choice of features and the overall design of the classifier [36].

#### 3.3.2.1 Bayes Basics Theorem

Let X is a data sample (evidence): class label is unknown and let H be a hypothesis that X belongs to class C. Classification is to determine P (H|X), (posteriori probability), the probability that the hypothesis holds given the observed data sample X.

P(H) (prior probability) is the initial probability or the prior probability of each class based on the training tuples and P(X|H) (likelihood) is the probability of observing the sample X, given that the hypothesis holds or conditional probabilities of attributes value for each class, P(X):

$$P(H \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H)P(H)}{P(\mathbf{X})} = P(\mathbf{X} \mid H) \times P(H) / P(\mathbf{X}) \qquad (3.4)$$

probability that sample data is observed and hence from the given training data X, posteriori probability of a hypothesis H, P(H|X), follows the Bayes' theorem [12].

Informally, this can be written as posteriori = likelihood x prior/evidence

### 3.3.2.2 Naïve Bayes Classifier

Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-dimensional attribute vector X = ($x_1$, $x_2$… $x_n$). Suppose there are m classes $C_1$, $C_2$… $C_m$. Then predict the class label of a tuple using Naïve Bayesian classification and the classification is to derive the maximum posteriori, i.e., the maximal P($C_i$|**X**).

This can be derived from Bayes' theorem

$$P(C_i \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C_i)P(C_i)}{P(\mathbf{X})} \qquad (3.5)$$

Since P(X) is constant for all classes, only

$$P(C_i \mid \mathbf{X}) = P(\mathbf{X} \mid C_i)P(C_i) \qquad (3.6)$$

needs to be maximized

Therefore, one can predicts X belongs to specific class if and only if the probability P ($C_i$|X) is the highest among all the P ($C_k$|X) for all the *k* classes.

A simplified assumption states that attributes are conditionally independent (i.e., no dependence relation between attributes) and yields:

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times ... \times P(x_n \mid C_i) \qquad (3.7)$$

Therefore, due to its lower error rate of predictive accuracy and capacity to provide a standard of optimal decision making, Naïve Bayes was implemented in this research experimentation.

### 3.3.3 The Weka tool

In this research work, Weka 3.6.software which is developed at the University of Waikato in New Zealand was used. This software is available at *www.cs.waikato.ac.nz/ml/Weka* site [11]. Weka tool is open-source data mining software in Java with a number of collections of algorithms for data mining tasks, including data preprocessing, association mining, classification, clustering, and visualization [47]. The tool has graphical user interface which consists of buttons and menu commands and panels on its interface, where each panel is used to perform different tasks. The initiation point in the tool after Weka has been run is graphical user interface with supporting tools. Next steps is selecting the application in the menu bar to process data mining tasks (explorer, experiment, knowledge Flow and simple CLI) by double clicking one of the options available on menu bar in order to perform the intended tasks. In order to start data mining tasks in the Weka, it is necessary to open a dataset from where it is saved and import to Weka tool. Once the explorer window is chosen and the application data file is imported, different techniques in the menu bar become active and one can perform different tasks according to proposed data mining objectives.

## 3.4 Methods of Training and Testing

The classifiers were evaluated by cross-validation using the number of folds. K-fold is a natural number used to check the performance of the model through k–times. K-Fold is appropriate whether the size of the data is very large or not. This is because of its extensive tests on numerous datasets with different learning schemes. It is also suggested that in k-fold algorithm, '10' is about the right number of folds to get the best estimate of error [14].

In 10-folds cross validation, the learning scheme or dataset is randomly reordered and then, split into 'n' folds of equal size. In each iteration, one fold is used for testing and the other n-1 folds are used for training the classifier. The test results are collected and averaged over all folds giving the estimate of accuracy for cross-validation. Therefore, k-folds minimize the bias effects by random sampling of the training and holdout data samples through repeating the experiments

ten times. To meet the intention of the research work, 70% was used for training purpose and the remaining 30% for validation of classifier accuracy.

## 3.5 Methods of Analysis and Evaluation of System Performance

After accomplishing model creation, comparing predictive accuracy of the classifiers for unknown tuples is often helpful to evaluate the performance of predictive modeling. It tells us how frequently instances of particular classes are correctly classified as actual class or misclassified as some other classes.

### 3.5.1 Confusion matrix

Confusion matrix is useful tool for analyzing how well classifier recognized the classes. It is body of table with m by m (row and column) matrix the row corresponds to correct classification and the column corresponds to the predicted classifications. An entry, CMi,j in the first m rows and m columns indicate the number of tuples of class that were labeled by the classifier as class j [28]. For a classifier to have good accuracy, ideally, most of the tuples would be represented along the diagonal of the confusion matrix with the rest of the entries being closed to zero [11].
In confusion matrix, there are classifier evaluation metrics like accuracy, error rate, sensitivity and specificity, precision, recall, and F-measure. Table 3.1 shows two class classification result simple confusion matrix which contains both predicted and actual classes.

**Table 3.1 Confusion Matrix with Two Classes Classification Result**

| | PREDICTED CLASS | | |
|---|---|---|---|
| ACTUAL CLASS | | Class=Yes | Class=No |
| | Class=Yes | a (TP) | b (FP) |
| | Class=No | c (FP) | d (TP) |

Key: TN= True Negative TP =True Positive FN =False Negative FP =False Positive

Here are some of performance evaluation computational techniques on confusion matrix that are used in this study. Accuracy is the first one which is widely used to check the performance of the model. It is the percentage of test set tuples that are correctly classified [48].

**Accuracy** = TP+TN= (TP + TN)/ (TP+TN+FP+FN) (3.8)

The performance of the model enables it to classify the positive cases correctly is sensitivity. It is defined as the probability of having a positive test result among those with a positive diagnosis for the disease or true Positive recognition rate [48].

**True Positive Rate (sensitivity)** = TP / (TP+ FN) (3.9)

The performance of the model to classify the negative cases is specificity. It is defined as the probability of having a negative test result among those with a negative diagnosis for the disease or true negative recognition rate:

**True Negative Rate** (specificity) or Recall for False class **=** TN / TN + FP (3.10)

Recall is what percent of positive tuples the classifier labeled as positive for both True and False classes (alive and died). Another detailed performance measure for the classifier is precision which measures what percent of tuples that the classifier labeled as positive are actually positive:

**Precision** = TP/ TP+FP----------For True Class (3.11)

**Precision** = TN/TN+FN----------For False Class (3.12)

Finally, the F measure is the inverse relationship between precision & recall ($F_1$ or F-score**)**: harmonic mean of precision and recall. It is the point to conclude that the precision and recall of the model are significantly balanced [48].

**F-Measure**=2xPrecision*recall/ (Precision + Recall) (3.13)

Error rate of the classifier is to determine how much percent error is committed by the model which is usually computed as the difference of one and accuracy. This is mostly appropriate if interpreted for classes with equal data distributions. Otherwise, it is recommended to test the model performance using ROC curve analysis.

### 3.5.2 Receiver Operating Characteristic (ROC) Curve

A large number of intelligent medical systems (including medical expert systems, neural networks, classifiers, knowledge discovery and data mining systems) showed great progress and they are being developed, practically to aid clinician and to improve patient care in areas such as diagnosis, prognosis, decision support and screening. To test which classifier is highly significant for a given subject is determined by ROC analysis and it becoming widely used tool in medical tests evaluation [49].

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified [50]. For example, it can be used to classify adults those who alive and died correctly based on their previous history. The following Figure 3.3 shows the performance of classifier B; that it has the maximum area under curve [27]

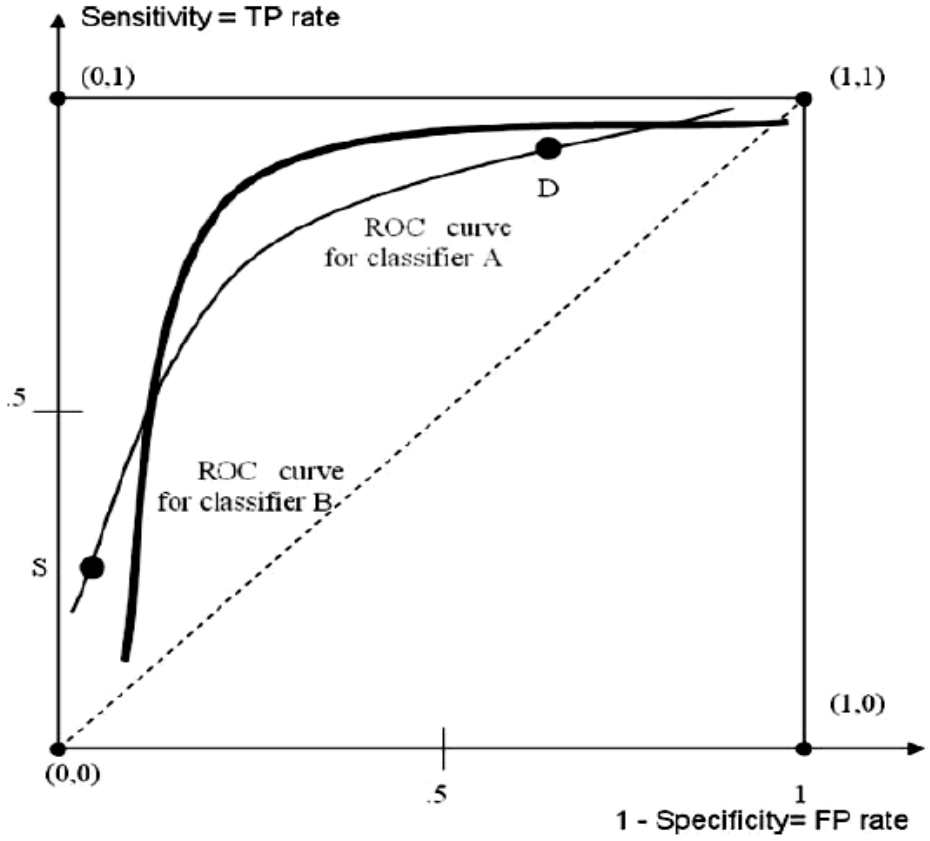


**Figure 3.3 Examples for ROC curve [20]**

ROC curve is useful visual tool for comparing classification models. It shows the trade-off between the true positive rate (proportion of positive tuples that are correctly identified) and the false-positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model [11, 27]. It is performed by drawing curve in two dimensional spaces by representing vertical axis for true-positive rate and the horizontal axis for false-positive rate [11]. In ROC curve, plotting starts at the bottom left-hand corner where the true positive rate and false-positive rate are zero. To plot an ROC curve for a given classification model, one need to rank the test tuples in decreasing order.

To assess the accuracy of a model, one can measure the area under the curve which is a portion of the area of the unit square and its value is ranged from 0-1. It is assumed that increasing numbers on the scale represents that the subject belongs to one category while decreasing numbers on the scale represent the increasing belief that the subject belongs to the other category [42]. Thus, from the ROC curve, the closer the ROC curve of a model is to the diagonal line, the less accurate the model is closer to the area of 0.5.

**Table 3.2 Performance Measures of ROC Area**

| ROC Area | Performance |
|---|---|
| 0.9-1.0 | Excellent(A) |
| 0.8-0.9 | Good (B) |
| 0.7-0.8 | Fair (C) |
| 0.6-0.7 | Poor (D) |
| 0.5-0.6 | Fail(F) |

The model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger evidence for a positive actual state (1.00) [11, 27, and 50]. For example, in Figure 3.3, classifier A performs better than B.

By using ROC analysis one can identify predictors in order to find the one with optimal characteristics and their associated cut-points. Therefore, sensitivity, specificity, precision, F-measure, and ROC area were taken in to account when the classifier performance is evaluated.

# CHAPTER FOUR

# BUSINESS UNDERSTANING AND PREPARATION OF BRHP DATA

To alleviate health care planning and strategies in rural Ethiopia in focus of adult health, in-depth critical assessment (domain problem understanding) is done to select features for constructing adult mortality predictive model. Understanding the problem is the most important step out of the six basic literature steps of Hybrid methodology [38]. Because the success of all the other steps depending on to what extent the problem is clear and the dataset selected for mining is related to the business area.

## 4.1 Problem Domain Understanding

Business understanding were carried out on pertinent issues giving due emphasis on the following parameters [27]. Firstly, working closely with domain experts to define the problem and determine the goal; secondly, identifying key people and learning about current solutions to the problem; thirdly, learning domain-specific terminology and finally, translating it into data mining goals and using the selected algorithms to derive knowledge. In this study we attempted to understand the problem area by dividing it in to major tasks and accordingly to identify what are the inputs used in the area? How inputs are processed? and what outputs area expected in return?

### 4.1.1 Task-1: Health promotion

It is a general non-specific intervention that enhances health of the individual in order to protect oneself from harm that leads to death; this is usually attained through improving socioeconomic and education status, provision of adequate water, clothing, environmental health and community development [9].

**Education**

It is a component of human development and vital to the all-rounded effort to eradicate devastating outcome on the society such as death, poverty, hunger etc. It is becoming a key development goal and considered as central to lift up the individual from unhealthy situation; to

lead sustainably in healthy and well being conditions. The MDGs have set quantitative targets for the reduction of death that related with lack of education through improvements in health.

Especial focuses on some indicators have given to track progress in achieving education in community and to harmonize the intended pledges on the millennium summit. These are, net enrollment ratio in primary education, proportion of pupils starting from grade 1 who reach grade 5, and literacy rate of adult particularly age 15-24 year olds [22]. Thus, in adult mortality reduction, efforts are being made by considering education as a central role to achieve the desired goal by increasing the literacy status of the individual since mortality can be explained by education [51].

**Water supply and safety measures**

In relation to water source of the individual/community, includes ensuring the availability and safety of domestic water supply are another area that need emphasis in adult mortality reduction [52]. In this sub section, creating individual/community awareness and sensitization on safety of water supply and identifying water sources for communities are important activities to maximize adult health through assessing and taking the corrective measures on the contaminated water using practice. Proper water handling at household level, inspection of water sources and storage including those of public water distribution sources are important areas to mitigate the health problem encountered with water sources. In health promotion strategies, ensuring the accessibility of the adequate water supply is emphasized to overcome death related with water borne in a given population.

Therefore, availability of safe water in the home with in fifteen minutes walking distance to whole population is considered as crucial policy approach to minimize the negative consequence on individual/community that resulted from water source [52].

**Health priorities**

To address the conditions that prevail health problem in individual/ community, setting priority is vital notion. Because it is impossible to address the entire problem at the same time due to the scarcity of resources. Here are some priorities given in health promotion in preventive and promotive focus [53]. These are development of preventive & promotive components of health

care, assurance of accessibility of health care for all segments of population, provision of health care for population on a scheme of payment for those who cannot afford to pay, advocacy efforts through promotion of participation of the private sector and nongovernmental organization in health care and special attention shall be given to the health needs of the family particularly women and children.

**Food hygiene and housing condition**

Hand washing, food storage, preservation, clean use of utensils, methods of food preservation, and identification of spoiled food are some of the important area when we talk about health promotion since these can independently affect the health of the individual [52]. A study conducted in Oslo [51] that aimed to examine the effect of housing conditions in childhood and adult cause-specific mortality states that sanitary conditions and economic deprivation appeared to be independently associated with mortality.

In general, health promotion is non-specific that directly or indirectly enhances the healthy condition of the individuals. Researches were conducted in various areas to identify the leading causes of adult death in order to tackle the source and to provide specific intervention. Some of the previous studies reveal that determinants of adult mortality broadly fall within a holistic framework of socio-economic, demographic and behavioral factors [48]. For example, here are some scholars' affirmation with regards to factors and adult death.

The three broad groups of causes of death were communicable and reproductive diseases, non-communicable diseases and injuries [25]. Huge burden for sub-Saharan Africa that results significant impact on health systems and social and family structure is HIV/AIDS [25]. Socioeconomic development has substantial role on adult health improvement and it needs emphasis to enhance adult health condition [18]

Therefore, studies conducted in different area to assess predictors of adult mortality state that there are predisposing factors for adult mortality. The following Table 4.1 shows some of the factors identified after business review.

**Table 4.1 Some of Adult Mortality Predictors Identified during Business Review**

| References | Attributes During Business Review | | | | | |
|---|---|---|---|---|---|---|
| [5, 24] | Period | Water source | Type of house | Sex | Malaria | Diarrhea |
| | Education status | Type of house | Age | Residential area | Tuberculosis | Other and unknown cause |
| [46] | Marital status | No education | No gainful occupation | Male sex | Living In rural lowlands | income |
| [1] | Educational | Sex | Type of residence | occupational | Income | Marital Status |
| | Smoking | Drinking alcohol | Drinking alcohol | Type of household | Type of fuel used | Chewing tobacco |
| [15] | Distance to Butajira | House ownership | Type of roof | Number of rooms | Timed | Distance to Butajira |
| | Literacy | Education | Source of water | Type of roof | Windows | House owner |
| [26,45] | Education | Rural lowlands | Male sex | Rural highlands | - | - |

### 4.1.2 Task -2: Disease prevention

Virtually, everything around us and everything we do affect our health; such as vital events like birth, death, marriage, and divorces, injuries, health related behavior like smoking, alcoholism, and social factors such as poverty are some of the examples [9]. Disease prevention is one of the important components of the public health promotion that aims to push back the precursors and risk factors of disease. In this regard, adverse outcome that is contributed by different factors are prevented. Therefore, attributes captured for disease prevention in business area are noted as socio-economic, demographic, behavioral and environmental factors which determine whether or not a disease developed [9]. In business area, awareness to keep from potentials of disease can save the adult from premature death.

### 4.1.3 Task-3: Advocacy efforts

In this regard, active supports on adult health were given to address the problem by collaborative actions i.e. governmental and non-governmental efforts. FMOH has strategies for improvement and mitigation of adverse impacts of factors beyond the health sectors. Among the strategies, adult illiteracy is considered as a threat of the health sector development program. This leads to poor health coverage of the adult and emphasis is being given to mobilize community based awareness through education [54].

In addition, there are also efforts to promote healthy behavioral life style in adult including reducing prevalence of daily smoking habit, heavy alcohol intake among adults, smoking and substance abused behavior [54]. In our country Ethiopia, there is literacy program that aimed primarily at literacy to enhance the skills and develop problem solving abilities of youth and adults; consequently that leads adult sustainably healthy and productive life [55].

Hence, comprehensive assessment were made in order to identify variables that role he/she die or alive. The intension is to enquire specific health problems in the study by quantifying particular issues that mainly affect adult health conditions and become the important issues to evaluate adult health care intervention. Therefore, heath promotion and disease prevention have a great attention in adult mortality so as to avoid factors which may cause disease and illness and if an individual is exposed to them and neglect to take early preventive action that may attribute to death.

In data mining, it is obvious that one need to select the variables that help to generate important rules that become applicable in the domain area. However, before this step exists, it is crucial to address the framework that illustrates the holistic area that need to be addressed in adult mortality. This is developed after reviewing the relevant literatures and discussion with domain experts.

Thus, association of adult mortality with different predictors were reviewed from different books, journals and idea from key peoples in the area with regard to adult mortality were gained to look the significant effects of predictors on adult health. To analyze the various factors on adult mortality, a conceptual framework which is recommended for such type of study by Saikia

et al. [1] was used by adding some important variables that determine adult health as depicted in Figure 4.1.

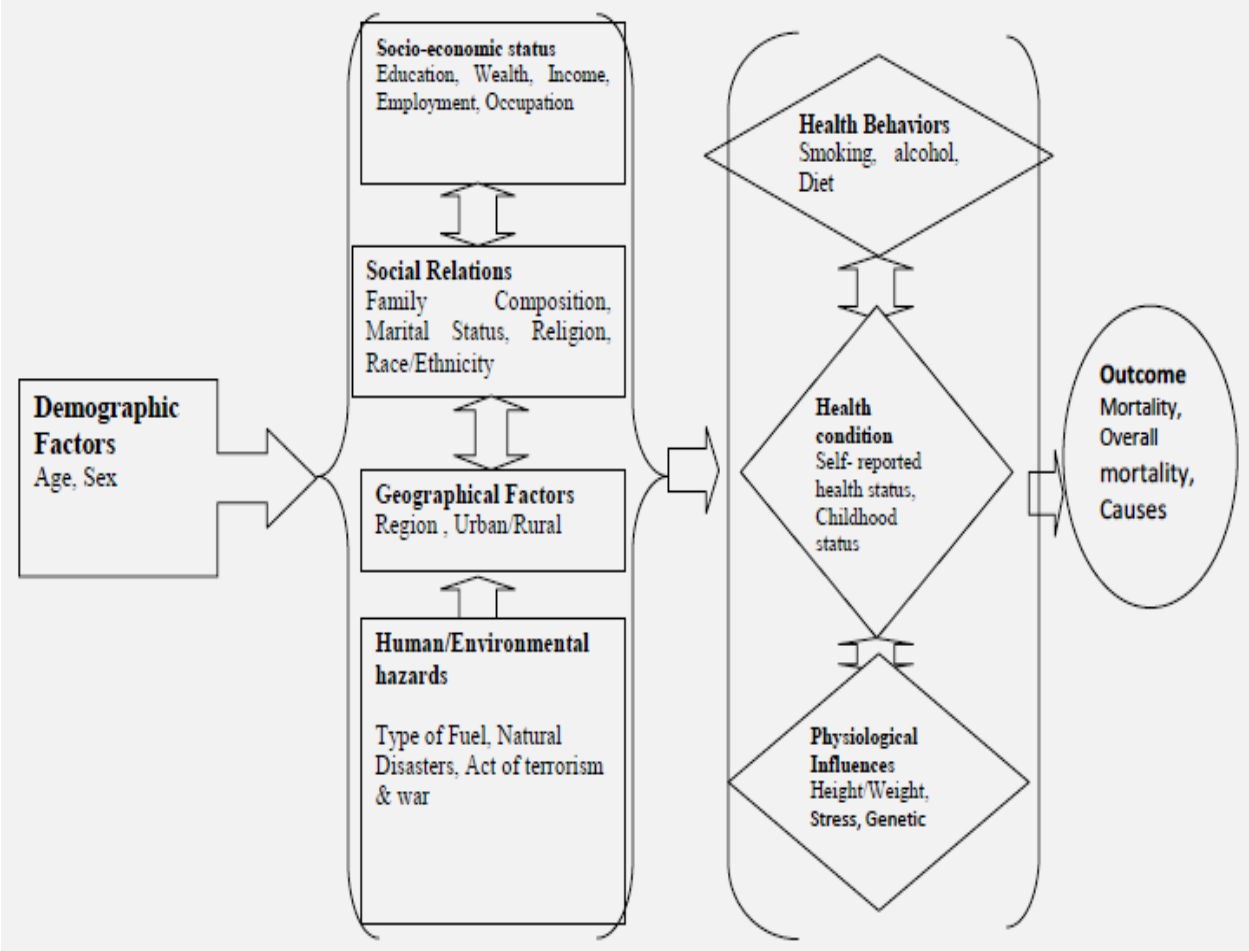


**Figure .4.1 Conceptual Framework for Factors Affecting Adult Mortality**

Potential interactions among predictors and adult death were assessed from various sources to put forth the predictors affecting adult mortality and to create a model that is useful for prediction. Attributes like living in rural area, sex of the individual, marital status, literacy, education status of the individual, water source, housing condition, and income of individual, exposure to mass media, type of fuel used for cooking, occupational status, behavioral life style like smoking, chewing tobacco, and drinking alcohol are some of adult mortality predictors identified from problem domain understanding.

## 4.2. Data Understanding

According to Cios et al. [27], data understanding phase mainly focuses on creating a target dataset with selected sets of variables that is relevant to discovery process. Without understanding the existing data, it is difficult to draw the target dataset from the origin since the world data is unclean and not appropriate at the source to run mining process.

Han and Kamber suggested [11] that attention should not be neglected to clean data for knowledge mining because the real world data is highly susceptible to noisy, inconsistency and incompleteness. Another idea that Han and Kamber added in the above suggestion is that the more the size of the data and the more multiple and heterogeneous source, the less the predictive performance of a model.

### 4.2.1 The raw data descriptions

The source data employed for this research purpose is BRHP data which was collected from 1987-2004 for the purpose of epidemiological surveillance in the rural Ethiopia based on different socio demographic and economic variables. This epidemiological surveillance data is collected from the nine randomly selected rural [4]*kebeles* known as peasants association and one urban kebele known as urban dwellers association by implementing probability-proportional-to-size technique [56]. The aim of this study is to create a model based on secondary data that was selected on the base of epidemiological surveillance in the district of Butajira. The entire attributes in the original dataset were not concerned for this experimentation. Thus, only relevant attributes were considered so as to achieve the objectives of the study.

From the whole dataset, only population with age 15-60 years was collected in order to satisfy the research goal in the study. Initially, the dataset is available in SPSS 16.0 format, and it contains a total of 320112 episodes (raw) and 34 attributes (column). Once a sample data was collected from the whole dataset on which feature selection and preprocessing is conducted, the next task is comprehensive assessment of distal and proximal determinants based on their significance level with respect to adult mortality. Table 4.2 describes BRHP raw data variables.

---

[4] *Kebele is the smallest administrative unit in Ethiopia.*

**Table 4.2 List of Variables in the Initial Dataset**

| Field | Field name | Type | Width | Measure ments | Descriptions |
|---|---|---|---|---|---|
| 01 | TTYREF | String | 6 | Nominal | 22-year record reference number |
| 02 | PA | String | 3 | Nominal | Peasant association code |
| 03 | ENVIR | String | 1 | Nominal | Environment of the household located |
| 04 | HOUSENO | String | 5 | Nominal | House number for residence |
| 05 | ID | String | 10 | Nominal | ID number |
| 06 | NAME | String | 20 | Nominal | Name |
| 07 | REL | String | 2 | Nominal | relationship with the head of house hold |
| 08 | SEX | String | 1 | Nominal | F = female, M = male |
| 09 | MID | String | 10 | Nominal | mother's id number |
| 10 | FID | String | 10 | Nominal | father's id number |
| 11 | MARITAL | String | 2 | Nominal | marital status during episode |
| 12 | SEREPI | Numeric | 2 | Nominal | serial no. of individual's episode |
| 13 | DBIRTH | Date | 12 | Scale | date of birth |
| 14 | RSTART | String | 2 | Nominal | reason for episode starting |
| 15 | DSTART | Date | 12 | Scale | date of episode starting |
| 16 | DEND | Date | 12 | Scale | date of episode ending |
| 17 | REND | Sting | 2 | Nominal | reason for episode ending |
| 18 | DDEATH | Date | 12 | Scale | date of death |
| 19 | TIMEX | Numeric | 4 | Scale | days of exposure during episode |
| 20 | CAUSE | String | 1 | Nominal | cause of death |
| 21 | RELIG | String | 2 | Nominal | individual's religion |
| 22 | LITER | String | 2 | Nominal | literacy during episode |
| 23 | EDUCATION | String | 2 | Nominal | educational status during episode |
| 24 | SOURCEW | String | 2 | Nominal | source of water during episode |
| 25 | ROOF | String | 2 | Nominal | type of roof during episode |
| 26 | WINDOWS | String | 2 | Nominal | windows in the house |
| 27 | RADIUS | Numeric | 2 | Nominal | radius of circular house in metres |
| 28 | ROOMS | Numeric | 2 | Nominal | number of rooms in house |
| 29 | HOUSEOWN | String | 2 | Nominal | house ownership: |
| 30 | OXEN | String | 2 | Nominal | number of oxen owned by family |
| 31 | TIMAD | Numeric | 2 | Scale | number of timad of land owned by family |
| 32 | LATITUDE | Numeric | 8 | Scale | latitude of household |
| 33 | LONGITUDE | Numeric | 8 | Scale | longitude of household |
| 34 | DISTHOSP | Numeric | 4 | Scale | distance to Butajira |

### 4.2.2 Attributes selection for knowledge discovery

This is to get a minimum set of best attributes for classification. Related attributes during problem understanding were selected for predictive model building. This is because considering the effects of predictors (independent variables) with predicted (outcome variable) whether they are linearly associated or not. As Han and Kamber suggestion [11], datasets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task; these possibly lead to do with domain expert to pick out some of the useful attributes by ignoring the irrelevant attributes. These in fact, maximize the quality of mining result as well as speed up the algorithm in mining process. But the question is, how a set of a good subset of the original attributes one can possibly find.

Han and Kamber suggested that [11] one can filter the best and worthy features using tests of statistical significance. This means that selecting best features based on their significance level potentially explain the class attribute (survive or die). If the attributes are dependent to each other, they may spoil the predictive accuracy of the model/classifier. Therefore, identifying how far the independent variables are correlated with each other is becoming a pre-requisite of mining process. Thus, attributes were checked through multi-colinearity analysis for their interdependences in order to make the outcome of interest free of confusion by which the outcome is really affected.

This can be checked using the tolerance and variance inflation factor; based on which one can decide the inter-predictors relationship. For example, it can be said that a tolerance value of one indicates that a variable is not correlated with others and a value of zero indicates that it is perfectly correlated and for VIF two shows a close correlation and one shows little correlation [55].

Based on this assumption, the colinearity diagnosis was checked and the result shows that there is no strong colinearity between the attributes since both tolerance and VIF are in the same region. Having bear in mind the above discussion to select the attributes from the original dataset, some of the variables removed from the sample dataset since they are less relevant for this particular study are house number, ID number, name, mother ID, date of death, date of birth, religion, latitude, longitude, father ID, ttyre (serial number of episode), and cause attribute.

Ignoring these attributes is based on business area and discussion with domain experts assuming that they are less important for adult mortality predictive modeling.

For example, religion is removed since it is a sensitive attribute to describe the study variable. There are some redundancies in the original dataset; for instance, the attribute distance from hospital includes all the information in the attributes of latitude and longitude. Therefore, no need of including the latitude and longitude as a risk factors for adult mortality since distance from hospital contains the relevant information. Though the cause is important factor for adult mortality, it was removed in this particular study because it deals only for died class not for alive. Descriptive data visualization was used in this section as presented in 4.2.

### 4.2.3 Descriptive data visualization

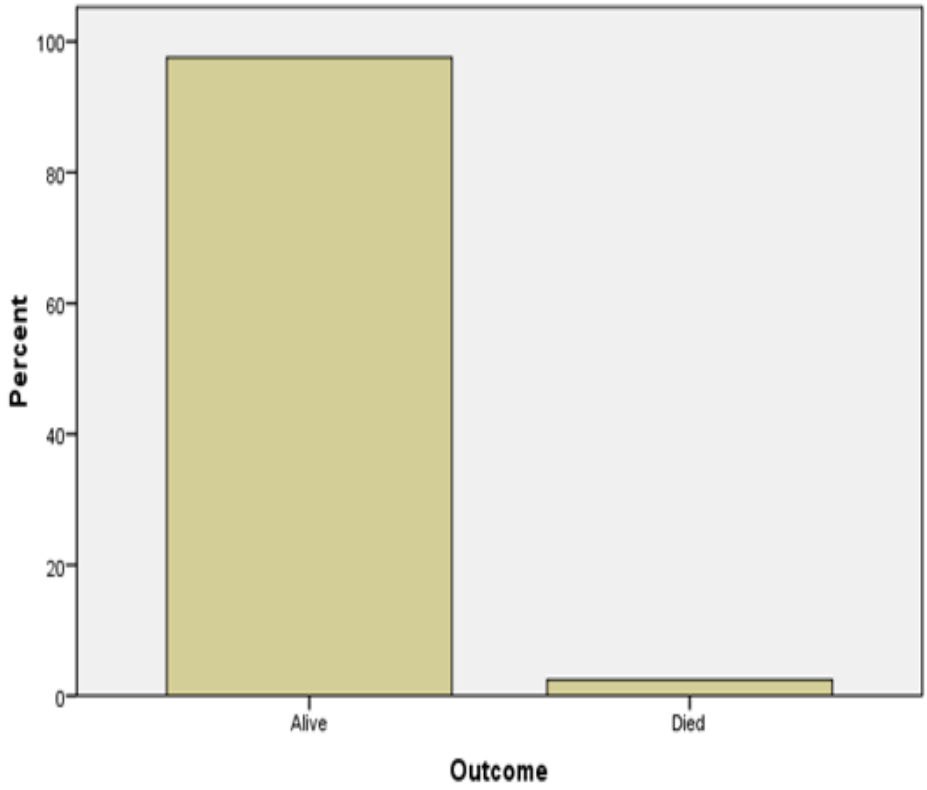


**Figure 4.2 Comparison between Died and Alive**

After selection of attributes from the entire dataset, statistical summary was done and all the attributes were checked for their significance effect on the outcome variable. With this regards, survival Cox regression analysis was also used in order to investigate significance of the

selected features on adult mortality. Table 4.3 below shows selected subset of attributes from the original dataset on the base of business review. The detail description of selected attributes is presented in annex 3.

**Table 4.3 Selected Subset Attributes from BRHP**

| Attributes | Type | Measurements | Descriptions | Remark | P-value |
|---|---|---|---|---|---|
| Residence | String | Nominal | Code of peasants villages | Living in rural is high risk than urban | P<0.05 |
| Environment | String | Nominal | Environment of the house hold is located | Climate | P<0.05 |
| Relation status | String | Nominal | Relationship with the head of house hold | Relation ship | P<0.05 |
| Sex | String | Nominal | F = female, M = male | Sex of the individual | P<0.05 |
| Marital status | String | Nominal | Marital status during episode | Marital status | P<0.05 |
| Time of exposue | Numeric | Scale | Days of exposure during episode | Days of exposure during episode | P<0.05 |
| Literacy | String | Nominal | Literacy during episode | Literacy | P<0.05 |
| Education status | String | Nominal | Educational status during episode | Education status | P<0.05 |
| Source of water | String | Nominal | Source of water during episode | Water source | P<0.05 |
| Roof | String | Nominal | Type of roof during episode | Types of roof | P<0.05 |
| Windows | String | Nominal | Windows in the house | Windows in house | P<0.05 |
| Radius | Numeric | Nominal | Radius of circular house in metres | Radius of circular House | P<0.05 |
| Rooms | Numeric | Nominal | Number of rooms in house | Rooms in house | P<0.05 |
| House own | String | Nominal | House ownership | Ownership of the house | P<0.05 |
| Oxen | String | Nominal | Number of oxen owned by family | Number of oxen | P<0.05 |
| Timad | Numeric | Scale | Number of timad | Number of timad | P<0.05 |
| Dstance to Butajira | Numeric | Scale | Distance to Butajira | Distance from Butajira | P<0.05 |
| Age | numeric | Scale | Age of the adult | Age of the individual | P<0.05 |
| Inmigration | String | Nominal | Moving inside the population | Moving in | P<0.05 |
| Outmigration | String | Nominal | Moving outside the population | Moving out | P<0.05 |

After selecting the features that determine adult health in the domain area, the next task is deciding on the number of sample dataset. As shown in Figure 4.2, out of the total, 97.6% were alive and 2.4% were died. The dataset of adult age 15-60 years for both classes (Alive or Died) were observed that they are imbalance.

According to Han and Kamber, if the data is skewed to one side just like our case, drawing representative sample is crucial to ensure each group is represented by equal number of instances and one can search for hidden knowledge on a sample instead of mining entire dataset [12, 32].

In this research work, sample data was drawn from both classes using stratified random sampling technique. To get the adequate amount of the sample from the demographic data that has been collected for the last 22 years, samples of records were obtained after stratification for ten peasant associations in the district. This was done by selecting the sample of cases from each peasant associations without replacement techniques available in SPSS statistical software. To select proportional samples of died adult from each peasant association; the ten villages were stratified into ten strata. From each strata using SPSS statistical tool, 25% sample records of died adult were selected randomly.

This process has resulted to a total of 2,715 died cases. In order to insure an equal representation of events for both classes (died and alive), 41,149 sample records about alive adults were also randomly selected from each of the ten-villages; a total of 43,864 selected cases from the ten villages based on both classes were obtained and used to build and test a model. The number of sample records selected and the procedures used to select a sample about both died and alive cases from the ten villages of the BRHP study area is summarized in the following Table 4.4.

**Table 4.4 Dataset Selection Procedures**

| PA code | Villages | Frequency | Alive | Selected | Died | Selected | Total |
|---|---|---|---|---|---|---|---|
| 005 | Mmeskan | 17667 | 17192 | 4298 | 475 | 238 | 4536 |
| 007 | Bati | 20620 | 19967 | 4992 | 653 | 327 | 5319 |
| 008 | Dobena | 21946 | 21035 | 5259 | 911 | 456 | 5715 |
| 011 | Bido | 10886 | 10576 | 2644 | 310 | 155 | 2799 |
| 04B | Dirama | 13028 | 12653 | 3164 | 375 | 188 | 3352 |
| 06A | Yeteker | 21228 | 20656 | 5164 | 572 | 286 | 5450 |
| 06B | Wrib | 24008 | 23381 | 5846 | 627 | 314 | 6160 |
| 09A | Mjarda | 14916 | 14488 | 3622 | 428 | 214 | 3836 |
| 09B | Hobe | 19069 | 18548 | 4637 | 521 | 261 | 4898 |
| K04 | Butajira | 61433 | 60882 | 1523 | 551 | 276 | 1799 |
| **Total** | | **224801** | **219378** | **41149** | **5423** | **2715** | **43,864** |

PA = peasant association code

## 4.3 Data Preprocessing for Mining

In this study, secondary data (BRHP) source was used. This data is organized by collecting facts from households located at Butajira rural health program that had been collected from 1986-2004 for the purpose of epidemiological surveillance in the rural Ethiopia. This contains records on socio demographic, sex, income, residences, sources of water, education, housing condition, environmental condition, causes etc. The 22 years data is used as a base for data collection, which is kept for the study in the school of public health, Addis Ababa University. It contains a total of 320,112 records of individuals registered in all the ten villages of the BRHP study area. Therefore, no further data collection mechanisms are employed since the collected data is ample enough to undertake the planned research. Creating a target dataset focusing on a subset of variables or data samples which is relevant on which discovery aimed to solve the problem was selected. Therefore, to increase mining process of the algorithms, understanding the information enfolded in the data was undertaken through solving the problems related with incompleteness, redundancy, and missing values that are undesirably incorporated with selected dataset.

### 4.3.1 Handling Missing Values

Missing attributes values in the data is most likely associated with unavailability of interesting information, lack of knowhow on the importance of data at the time of entry, misunderstanding of the data, the respondents him/her self may refuse to answer certain questions or they may not know the answer exactly or may answer in an unexpected manner. Susceptibility of data for noisy, in consistency and incompleteness is becoming dominant in real world; this is mainly related with its huge size and heterogeneous source of data [11]. Thus, it is important for the researcher to manage missing values efficiently. If the missing values are not handled properly, handling data missing values in the dataset in filtering to get quality data that substantially improves the overall quality of the patterns mined is very crucial. This contributes for the success of timely mining process.

As it is suggested [11], it is possible to ignore the tuple when the class label is missing especially for classification tasks. But, this method is not effective unless the tuple contains several attributes with missing values and if the attributes' effect is significant for outcome of interest. Here in our case there are such kinds of tuples with missing values which are represented by a '99'. These are timad, room, and radius as depicted in Table 4.5

**Table 4.5 Statistical Summary of Attributes with Missing Values**

| | | **Timad** | **Radius** | **Room** |
|---|---|---|---|---|
| **N** | **Valid** | 174736(77.7%) | 132804(59.1%) | 212560(94.6%) |
| | **Missing** | 50065(22.3%) | 91997(40.9%) | 12241(5.4%) |
| Total | | 224801(100%) | 224801(100%) | 224801(100%) |
| Mean | | 2.95 | 3.40 | 1.53 |
| Median | | 3.00 | 3.00 | 1.00 |
| Mode | | 2 | 3 | 1 |
| Std. Deviation | | 2.512 | .724 | 1.017 |
| Skewness | | 18.067 | .160 | 2.705 |

**Tale 4.6 Missing Values Handling Mechanisms**

| No | Attributes name | Attributes type | Missing | Substituted mean value |
|---|---|---|---|---|
| 1. | Timad | Numeric | 50065(22.3%) | 2.95 |
| 2. | Room | Numeric | 12241(5.4%) | 1.53 |
| 3. | Radius | Numeric | 91997(40.9%) | 3.40 |

After missing values have been detected, the next task is filling the missing values with their mean for all episodes belonging to the same class using statistical tool (SPSS) as depicted in Table 4.6

### 4.3.2 Handling outliers

After replacing missing values, the next task is detecting outliers since data can be over dispersed by various factors. Outliers are one of the factors that potentially hinder the knowledge in the data which do not comply with the general behavior or model of the data [11, 58].

According to Han and Kamber [11], a common rule of thumb for identifying suspected outliers is to single out values falling at least 1:5XIQR (Inter Quartile Range) above the third quartile or below the first quartiles which are upper and lower adjusted values respectively. In this case, the box plot which is a graphical representation of a dataset was used to develop a visual impression of location, spread, and the degree and direction of skewness as depicted in Figure 4.3.

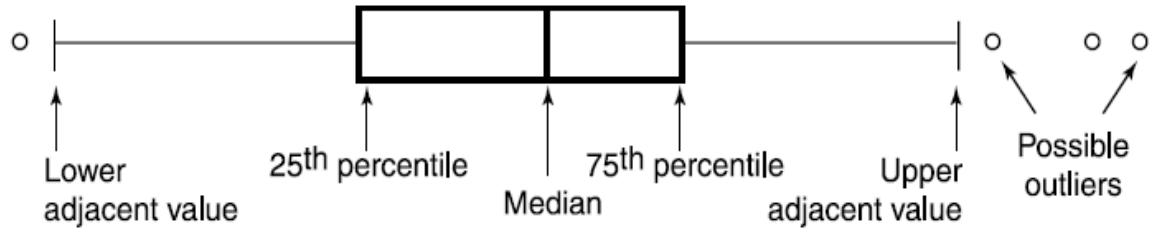


Figure 4.3 Box Plot to Detect Outliers

All points outside both lower and upper adjustment layers are represented by circles and considered to be outliers [58] for outlier detection was employed for this sub section. Upper adjustment = Q3+(1.5*IQR) and Lower adjusmtent =Q1-(1.5*IQR). Outliers in this section were detected based on the statistical analysis of Table 4.7.

**Table 4.7 Table for Twenty Fifth and Sevety Fifth Percentile**

| | | Median | Mode | Min. | Max. | Percentiles | | |
|---|---|---|---|---|---|---|---|---|
| Missing | Mean | | | | | 25 | 50 | 75 |
| ROOMS | 1.5271 | 1.0000 | 1.00 | 1.00 | 8.00 | 1.0000 | 1.0000 | 2.0000 |
| TIMAD | 2.9537 | 3.0000 | 2.00 | 0.00 | 98.00 | 2.0000 | 3.0000 | 4.0000 |
| RADIUS | 3.4070 | 3.0000 | 3.00 | 2.00 | 10.00 | 3.0000 | 3.0000 | 4.0000 |
| DISHOP | 7.694 | 7.600 | 0.9 | .4 | 20.8 | 1.400 | 7.600 | 11.100 |

Therefore, considering the formula shown above, all the values above and below the adjustment was considered as outlier. Calculation used for outlier detection is presented in Annex 1. Once outliers are detected, all the values were replaced by their respective attribute series mean. The following Figure 4.4 indicates box plots used for visualize the outliers in the dataset. Age attribute is used in the plot to ilustrate it has no outliered value.

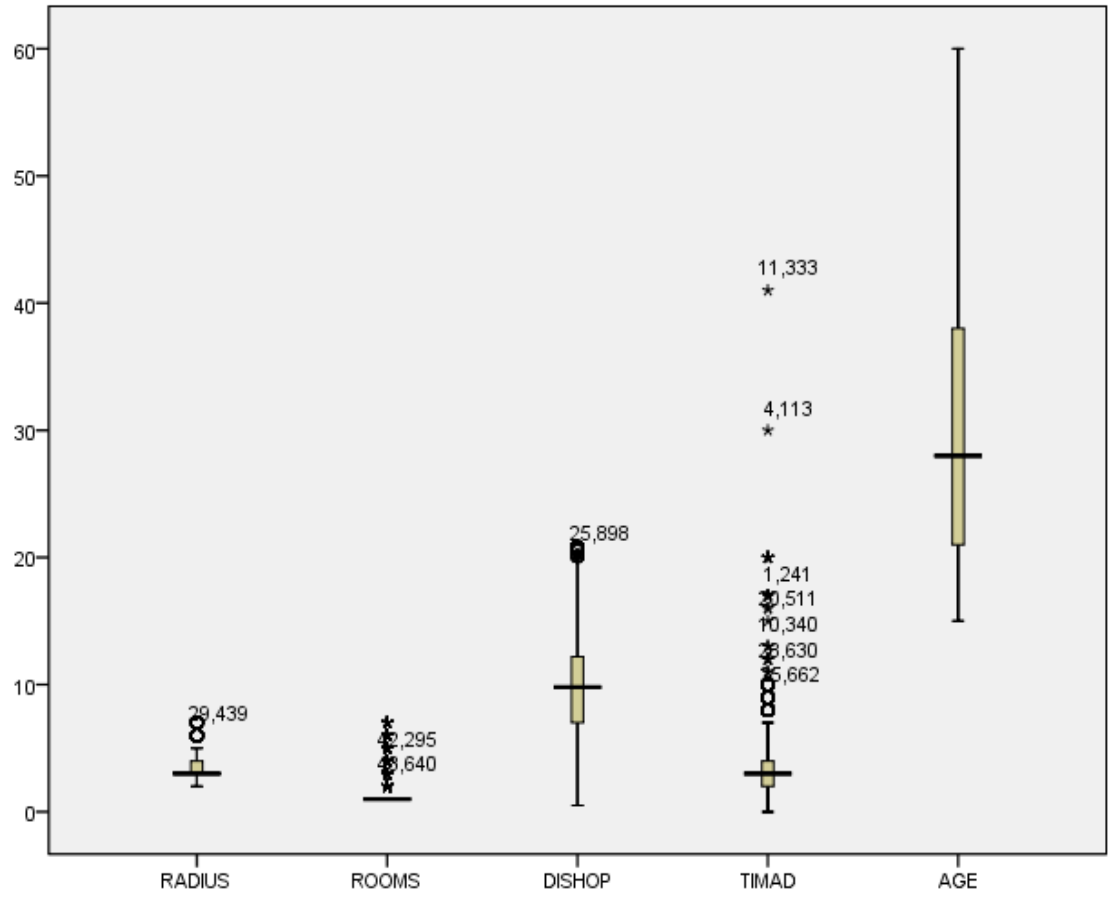


Figure 4.4: Box Plot used for Numeric attributes to Detect Outliers

### 4.3.3 Data decoding

Although the original dataset was organized in rows (episodes) and columns (attributes) using SPSS statistical software, it was not possible to import this dataset as it is into the data mining software (Weka). Therefore, data decoding is a necessary step in data cleaning process

Data decoding was done through allocating a new set of replacement values for a set of values in a given attribute such that each old value can be identified with one of the new values. Specially, if numbers represent a unique concepts and non-value with in continuous range of some quantity it needs conversion in to symbols [15]. Here in our case there are attributes with numeric codes (PA). For the purpose of limiting the confusion during model creation such types of numeric codes were converted to their respective symbols. The alphanumeric code given for the peasant association for nine rural kebeles and one urban kebele was regrouped in to their actual values as depicted in Table 4.8.

**Table 4. 8 Peasant association Attribute Transformation**

| Peasant association code | New value represented |
|---|---|
| 005 | Meskan |
| 007 | Bati |
| 008 | Dobena |
| 011 | Bido |
| 04B | Dirama |
| 06A | Yeteker |
| 06B | Wrib |
| 09A | Mjarda |
| 09B | Hobe |
| K04 | Butajira area 04 |

### 4.3.4 Attributes transformation

Data transformation has momentous effect on data mining since it helps to fix the problem of missing values in the data and brings information to the surface by creating new features to represent trends and other ratios [38]. In this study, through data transformation, the following attributes are obtained.

**Age**: Considering some variable as relevant for the adult mortality predictive modeling, they were created from the original dataset. Age attribute was derived from date of birth and from the last census enumeration by using statistical tool (SPSS). According to the objective of the study, individuals ages 15-60 were selected for this research work. This helps to look the problem existence in different age groups and helps to bring a solution depending on the clue from the study.

**Residence:** It was created from the original dataset of PA (peasant association code) attribute. This is considering the business area suggestion. Business tells that residence as one of the influential factors in adult mortality. In the existing dataset there is PA attribute which is represented by alphanumeric number codes. These codes were regrouped in to two areas, i.e. the nine rural kebles as a rural and the K04 as urban kebele; then we have residence as a factor for adult mortality

**Inmigration:** This attribute was created from reason for episode starting. Considering as important parameter of reason for episode start inmigration attribute ws created.

**Outmigration**: This attribute was created from reason for end. Considering Outmigration as important parameter in domain area, it was created from reason for end. Using them as they are in the original dataset, leads to overlapping of the features(reason for start and reason for end) such that it is impossible to mine the knowledge and become obstacle in pattern extraction.

**Status attribute:** These are derived from rend (reason for end) since they are variables intended in the study (died, alive).

#### 4.3.5 Weka understandable format

After all, the next important issue was importing the selected dataset from SPSS document format into Ms-Excel format in order to create Weka understandable format (arff and csv) for experimentation. Table 4.9 presets comparison of the original dataset and the refined target dataset.

**Table 4.9 Summary of Original and Target Datasets**

| **Parameters** | **Original dataset** | **Target dataset** | | |
|---|---|---|---|---|
| Fields | 34 | 27 | | |
| Total Number Of Records | 320,112 | 62,869 | | |
| File Format | SPSS 16.0 | .xls | .csv | .arf |
| Size of Data | 84.3MB | 12.1MB | 3.05MB | 4.83MB |

# CHAPTER FIVE

# EXPERIMENTATION

## 5.1 Overview of Experimentation

In this study different experiments were conducted using various data mining methods to derive knowledge from preprocessed data to predict unseen episodes of adults.

According to the methodology of this study after preparation of the data, the next task is the mining process. As it has been stated in the previous sections, from the both classes, 43,864 cases consists of 41149(97.6%) alive and 2715(2.4 %) died dataset were applied for experimentation.

To avoid the effect of data imbalance on the model created, Weka based SMOTE technique is applied. This is automatic operation where minority classes are over sampled by generating synthetic examples of minority class and adding them to the dataset to make the target attribute balanced [59]. According to Larose [60], if the class attribute is imbalance, this condition further need balancing by different techniques. Unless the classes are proportional, the classification will be skewed in dominant classes. Consequently, the new predicted instances will also fall in the dominant classes erroneously unless the classes' proportionality is considered.

Therefore, assuming such kind of problem in our case, SMOTE technique on the dataset selected for both training and testing have been applied using Weka data mining tool. Thus, the classification accuracy of the minority class become increased in the SMOTE technique for certain level i.e. the total of 62,869 cases (41,149 alive and 21,720 died) were provided and the subsequent experimentations were conducted based on this sample dataset. Figure 5.1 shows the original dataset and the balanced one after applying the SMOTE technique.


(a) Before SMOTE (b) After SMOTE

**Figure 5.1 Side by side Review of the class variable using SMOTE**

For experimentation, different algorithms were employed considering different parameters for model building such as pruning, unpruning and testing model performance with selected attributes and all attributes in both training and testing phases as depicted in Table 5.1.

**Table 5.1 Experiments and Scenarios**

| Experiments(1-3) | Scenarios(1-6) |
|---|---|
| J48 Unpruned Tree Model Generation | J48 unpruned with all attributes |
| | J48 unpruned with selected attributes |
| J48 Pruned Tree Model Generation | J48 pruned tree model with all attributes |
| | J48 pruned tree model with selected attributes |
| Naïve Bayes Classifier | Naïve Bayes with all attributes |
| | Naïve Bayes with selected attributes |

To build predictive model, 62,869 instances and 21 attributes were used through using both Naïve Bayes classifier and decision tree algorithm. The models generated with all attributes were compared with models with selected attributes.

To evaluate the performance of the models, the researcher used ten various number of percentage split starting from 10 to 99 since it ranges from 1 to 100 excluding the two extreme borders**.** In this scenario, all the results were closely equal but the only difference observed was variation of the execution period. The lowest execution time taken was observed in 70 percentages split.

Therefore, in this study, the learning procedure was executed by using 10-folds cross validation on different training and testing sets with 70 percentage split in the three experiments and six scenarios. Moreover, the performance of the models in this study was evaluated using the standard metrics of the accuracy, precision, sensitivity, specificity and the ROC area using the predictive classification table, confusion matrix.

Finally, the J48 pruned tree model with an accurate result of prediction 97.127% is selected and rule tracing and further analysis was employed. In this section, the importance of attributes in adult mortality predictive modeling were checked by maximum gain ratio using Weka attribute ranking optimal attributes as depicted in Table 5.

**Table 5.2 Attributes Rank with Information Gain**

| **Name Of Attributes** | **Information gain** | **Name Of Attributes** | **Information gain** |
|---|---|---|---|
| Dishop | 0.6733 | Roof | 0.1203 |
| Age | 0.6155 | Windows | 0.1084 |
| Timad | 0.554 | Source of Water | 0.0905 |
| Radius | 0.4863 | Residence | 0.0877 |
| Rooms | 0.3224 | Relation | 0.0819 |
| Time of Exposure | 0.3218 | Outmigration | 0.0787 |
| Education | 0.2587 | Environment | 0.0723 |
| Marital Status | 0.2339 | Inmigration | 0.0675 |
| Literacy | 0.1855 | House ownership | 0.0555 |
| Oxen | 0.1317 | Sex | 0.0265 |

## 5.2 Selecting and Evaluating the Attributes

Attributes selection involves searching through all possible combinations of attributes in the data to find which subset of attributes work best for prediction. To do this, determining CfSubsetEval method was used to assign a worth to each subset of attribute by searching ranker style in the Weka. With this regards, attributes selected using best first techniques in Weka are sex, literacy, education, distance form hospital, age, radius, timad, rooms, in migration and out migration.

During experimentations, all the default parameters that already set in the Weka were used except that the unpruned 'False' was changed to unpruned 'True' during J48 unpruned tree generation in order to experiment the model performance without pruning the tree.

**Table 5.3 J48 Classifier Parameter Options**

| Parameters | Descriptions | Parameter type |
|---|---|---|
| binarySplit | When to use binary split on nominal attributes when building the trees. | Boolean |
| confidenceFactor | The confidence factor used for pruning (smaller values incur more pruning). | Numeric |
| Debug | If set to true, classifier may out put additional info to the console. | Boolean |
| minNumObj | The minimum number of instances per leaf. | Numeric |
| NumFolds | Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree. | Numeric |
| reducedEroorPruning | Whether reduced-error pruning is used instead of C.4.5 pruning. | Boolean |
| saveInstanceData | Whether to save the training data for visualization. | Boolean |
| Seed | The seed used for randomizing the data when reduced-error pruning is used. | Numeric |
| subtreeRaising | Whether to consider the sub tree raising operation when pruning. | Boolean |
| Unpruned | Whether pruning is performed. | Boolean |
| usedLaplace | Whether counts at leaves are smoothed based on Laplace. | Boolean |

## 5.3 Model Building

To build the model, three different experiments were conducted using J48 decision tree and Naïve Bayes classifier. The intention here is to investigate the effect of attribute selection on classification accuracy as well as model complexity and decision tree size on both pruned and unpruned J48 tree classifiers. The second algorithm, Naïve Bayes classifier was evaluated on model performance with all and selected attributes. Before conducting different experiments, sample dataset was learnt by assigning different dataset for each type of set in model creation and model usage. To get the best performance of the model, the researcher conducted using different cross validation values on both training and testing schemes as depicted in Table 5.4.

**Table 5.4 Performance** of the Classifier for Different K-Values

| performance measure | Number of cross validation folds (K values) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Accuracy (%)** | 96.8 57 | 96.9 174 | 97.1 003 | 97.1 274 | 97.1 083 | 97.1 035 | 97.0 781 | 97.1 894 | 97.17 83 |

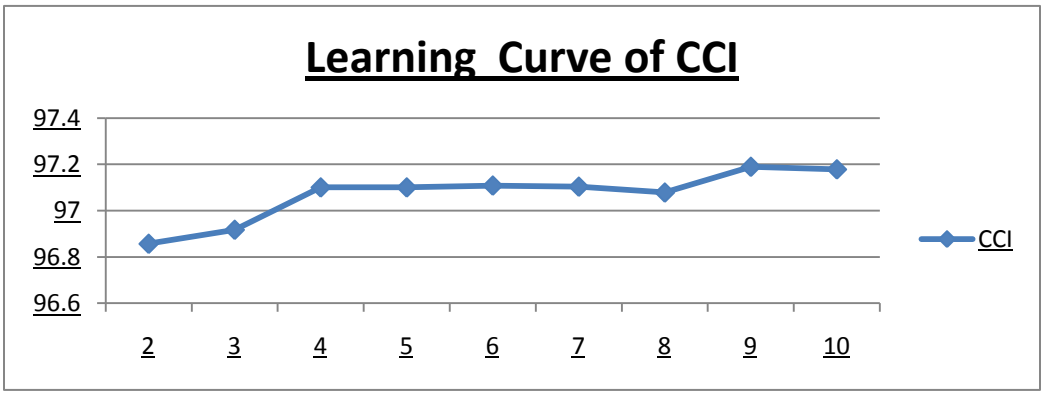


**Figure 5.2 Learning Progress Curve**

As indicated in Figure 5.2, though the performance variations among different k values are minimal i.e. nearly 97% successes, a bit higher performance was observed in 10 fold k-folds with its minimal error rate to classify the instances in wrong classes.

The following Table 5.5 indicates the performance summary of the experimental results for all experiments.

**Table 5.5 Performance Summary of the Models**

| Experiments of the detailed accuracy by class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Accuracy** | TP rate | **TN rate** | **precision** | **Recall** | **F-measure** | **ROC area** | **size of tree** | **# of leaves** |
| **Scenario1** | 96.9 % | 97.8% | 95.2% | 0.969 | 0.969 | 0.969 | 0.973 | 3984 | 3046 |
| **Scenario2** | 93.6 % | 95.6% | 89.8% | 0.936 | 0.936 | 0.936 | 0.96 | 3012 | 1573 |
| **Scenario3** | 97.2% | 98.5 % | 94.6% | 0.972 | 0.972 | 0.972 | 0.983 | 1031 | 715 |
| **Scenario4** | 93.7 % | 96% | 89.3% | 0.937 | 0.937 | 0.937 | 0.964 | 2043 | 1054 |
| **Scenario5** | 95.7% | 98.1% | 91.3% | 0.957 | 0.957 | 0.957 | 0.983 | - | - |
| **Scenario6** | 94.7% | 97.1% | 90% | 0.946 | 0.947 | 0.946 | 0.976 | - | - |

Various results have been executed using J48 unpruned tree, J48 pruned tree, and Naïve Bayes models. The first experiment shows that J48 unpruned decision tree algorithm with all attributes is capable in adult mortality predicting with performance of 96.9%. The result of J48 unpruned decision tree classifier with selecting attributes also has significant effect on classification and prediction accuracy of adult mortality (93.6%). The second experiment indicates that J48 pruned decision tree algorithm with all attributes is highly competent of 97.2 % in accurate adult mortality prediction. J48 prune decision tree classifier with selecting attributes also has promising effect on predicting adult mortality which performs 93.7%. Finally, the third experiment was designed to evaluate the performance of the Naïve Bayes algorithm in predicting adult mortality. The result indicates that the classifier performance with all attributes is slightly higher than classifier performance with selected attributes; the respective figure is 95.734% and 94.6571 %. The following Figure 5.3 shows variations of accuracy among models.

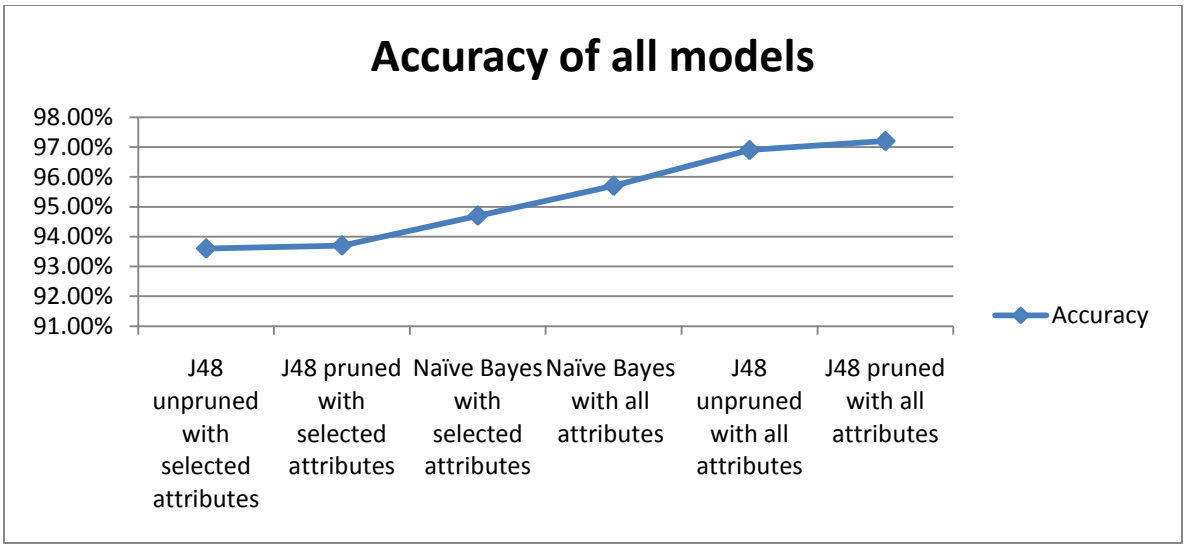


**Figure 5.3 Accuracy of All Models in the Experiments**

Moreover, the predictive performances of the models were also compared using different performance measures as testing criteria of the models.

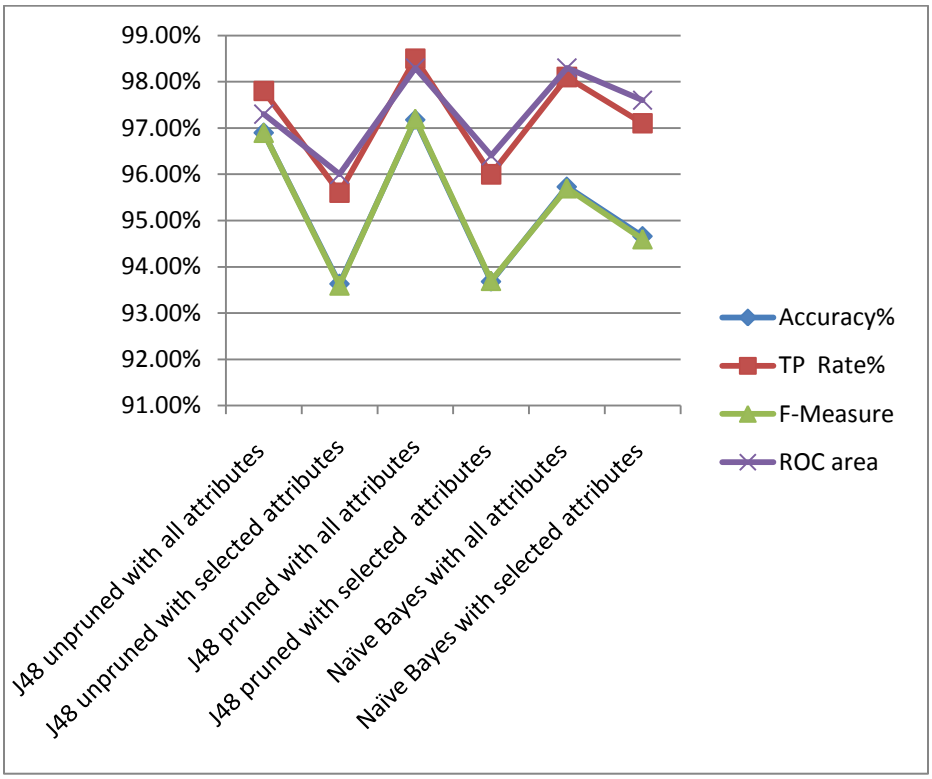


**Figure 5.4 Performance Comparison of the Models**

These are TP rate, TN rate, precision, and F-measure and ROC area. With regards to TP rate, J48 pruned tree score of 98.5% while the lowest score is observed in J48unpruned with selected

attributes (95.6%). The highest score of 95.2% TN rate in the experiment is registered by J48 unpruned tree model with all attributes followed by J48 pruned tree model with all attributes which registers 94.6%. Figure 5.4 shows the comparison of the performance measures.

In general, J48 pruned tree model, J48 unpruned tree and Naives Bayes are appeared with competent predictive performance for adult mortality. From all the scenarios experimented, all models reveal the better performance in predicting True positive cases or sensitivity (alive); than predictive performance of True negative case or specificity (died). This is the fact due to the model committing a bias to majority class over the minority class (alive and died) respectively.

The next task in testing the model to decide which one of the six models constitutes a better model/classifier of the BRHP data is evaluated using ROC analysis. This is Receiver Operating Characteristic analysis in which the curve the more to the upper left would indicate a better classifier [20]. Here in our case, the ROC area performance of the algorithms show that J48 pruned tree algorithm with all attributes and Naïve Bayes with all attributes scored the highest area of 0.983. The lowest ROC keeps account in J48 unpruned with selected attributes which is 0.96.

With regards of ROC area, a model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicates the stronger evidence for a positive actual state. Therefore, from the result, J48 pruned tree model and Naïve Bayes classifier have a good capability of 98.3% in adult mortality prediction. Based on the above assumptions, J48 pruned tree model with all attributes has an outstanding performance of 97.2% accurate prediction with 0.983 ROC areas.

Thus, J48 pruned tree model with all attributes is selected as the best model and its confusion matrix and ROC area are presented in the following session. The experimental outputs of others are presented in Annex 2.

### 5.3.1 Selected model performance and evaluations

During J48 pruned tree model generation, the effect of the attributes on the model performance was investigated. The full training set containing a total of 62,869 instances were used in all and selected attributes. In addition to above performance metrics used, in terms of tree size and number of leaves, relatively J48 pruned with all attributes is more understandable and less

complex to human than other others model generated. Therefore, the performance of J48 pruned tree classifier with all attributes gives a valuable information in predicting adult mortality as compared to models with all and selected attributes. The following Figure 5.5 shows the output of J48 pruned tree model.

```
Scheme:       Weka.classifiers.trees.J48 -C 0.25 -M 2
Attribute:          21
Test mode 10-fold cross-validation
J48 Pruned  Tree
Number of Leaves  :   715
Size of the tree :        1031
Time taken to build model: 5.99 seconds
Correctly Classified Instances       61095               97.1783 %
Incorrectly Classified Instances      1774               2.8217 %
Total Number of Instances           62869
Detailed Accuracy By Class
TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
0.985     0.054       0.972       0.985       0.979         0.983         Alive
0.946     0.015       0.971       0.946       0.959         0.983          Died
0.972      0.04       0.972       0.972       0.972         0.983         Weighted Avg.
= = = Confusion Matrix = = =
a     b   <-- classified as
40540   609 |     a = Alive
1165 20555 |     b = Died
```

**Figure 5.5 J48 Pruned Tree Model with All Attributes**

In J48 pruned tree predictive model, the predictive performance of the model is 97.2 % i.e. 61, 1095 instances were classified correctly while 1774 (2.9%) instances were wrongly misclassified to other class. The model classified 40540 instances as alive out of 41149 instances that in fact they are alive as tested on the test data or which are classified correctly in the class of alive. The remaining 609 instances were misclassified to another class as died actually they are alive.

The model classified 20555 instances as died out of 21720 instances that in fact died and wrongly classified 1165 instances to other class as alive while actually they had died. The model has a good performance in classifying the instances in True class (TP) than True negative class

(TN) (alive and died) with predictive performance of 98.5% and 94.6% respectively. Thus, it is possible to conclude that the model is in a good performance to classify True positive than True negative.

From the precision score of the model, the precision of this model for 'Alive' class is a bit higher than precision of 'Died' class (0.972 and 0.971) respectively. With an average precision of 97.2%, instances labeled as belonging for each class Yes and class No (alive, died ). From harmonic mean of precision and recall which is F-score, with value of 0.972, it can be concluded that the precision and recall of the model are significantly balanced.

When we come to the ROC curve of the selected model, the true positive case (sensitivity) and false positive case (specificity) are represented by vertical axis and horizontal axis, i.e. instances are predicted as alive actually they are alive and predicted as died actually they were died respectively. Figure 5.6 shows the ROC area of J48 pruned tree model with all attributes.

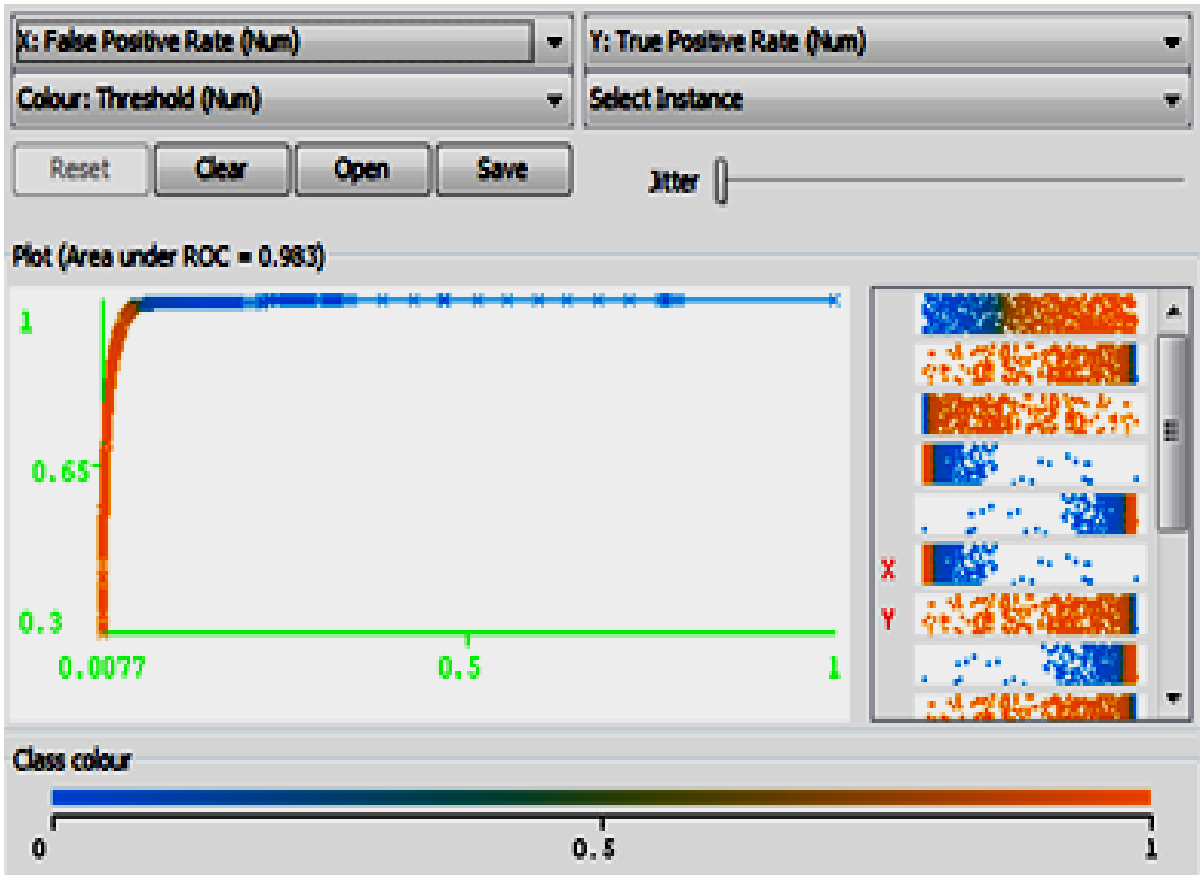


**Figure 5.6 ROC Area of J48 Pruned Model with all attributes**

In the above curve, Figure 4.6, initially it moves sharply up from zero showing that the model is better in detecting true positive than false positive. At the end, the curve trade off and become more horizontal showing that from the point where the curve starts to bend to onwards, false positivity outweighs true positivity i.e. the more the curve bend to the right, the more the false positivity rate and the less true positivity rate. The area under the model J48 pruned tree model is 0.983 which is closer to 1 showing that the class value yes gives ROC accuracy of 98.3%. In general, the performance measure of the selected model is depicted in Figure 5.7.

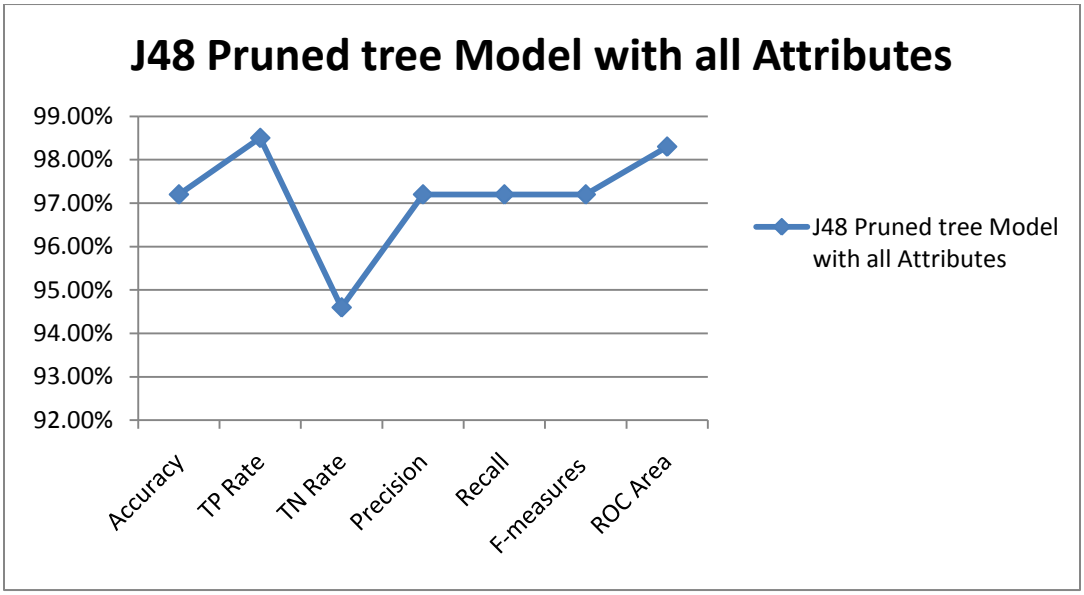


**Figure 5.7 Performace Measures the Selected Model**

## 5.4. Discussion

In this study, three experiments have been conducted using two data mining classification algorithms i.e. J48 algorithm and Naïve Bayes classifier in order to build a model that predicts adult mortality status. The goal in this study was to explore BRHP dataset in order to build the model that can predict adult mortality and to discover interesting patterns which are hidden in BRHP dataset, finally to notice the attributes that existed as predictors of adult mortality.

Wirth regards of effect of decision tree pruning, it is obvious that model with grown size of tree make the model difficult to understand and interpret by human as well as generating the rule become challenging. The experiment was done to reduce the complexity of the tree so as to make model more compact and understandable. Therefore, the models were experimented through

pruning and unpruning the tree on the training schemes. The result indicates that the effect of pruning is strong and number of leaves reduced from 3046 to 715 and size of the tree from 3984 to 1031. In pruned tree model, not only size of tree and leaves reduced but the accuracy of the model is improved from 96.9% to 97.2%. This is due to the fact, avoiding of branches that return unnecessary effects on the model

J48 pruned tree model was selected with its performance that could predict status of adult (alive, die) i.e. 97.2% accurate prediction with the respective concordant (True positive and True Negative) with the lower mean absolute error (0.0412) which measures the error between actual and predicted value and with high kappa statistic measures (0.9372); it is usually 1.0 which implies complete agreement. In this study, the model created using J48 pruned tree registers good performance and hence selected for further analysis/rule tracing. The first six most important parameters/attributes that determine adult mortality are education status, outmigration, inmigration, literacy, residence, and distance from the Butajira town. Some previous studies [26, 45, and 61] also ensure these attributes as predictors of adult death.

### 5.4.1 Rule Extraction

To make a decision tree model more human-readable each path from root to leaf can be transformed into an IF-THEN rule. If the condition is satisfied, the conclusion follows. The algorithm decision tree is the best known method for deriving rules from classification trees. Figure 5.8 shows the partial decision tree generated for BRHP dataset.

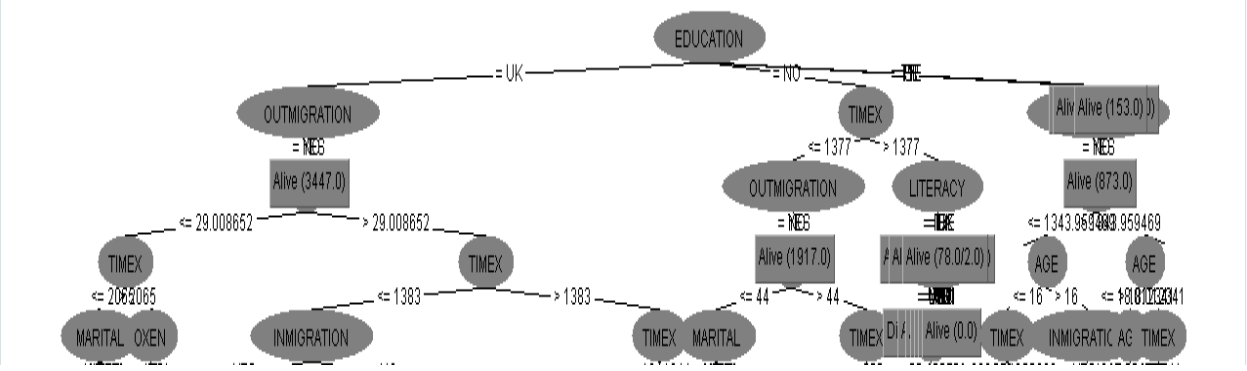


**Figure 5.8 Partial Decision Tree Generated For BRHP Dataset**

This is simply by traversing any given path from the root node to any leaf. The numbers in (parentheses) at the end of each leaf indicates the number of examples in the leaves. The number of misclassified examples would also be given, after a slash and hence it is possible to compute the success fraction (ratio) to estimate the level of confidence or likelihood of predictability of the class that tells how much the rule is strong.

From the entire models that were generated, J48 pruned tree model with all attributes is selected as the best model for rule generation. This is due to the rules provided by decision tree models can be easily assimilated by human without any difficulty.J48 pruned tree model with all attributes produced different rules. However, the researcher selected best rules that cover most of the data points in the study. The partial decision tree generated is presented in annex 4.

After the rule extraction, the researcher turns back to domain experts to discuss up on the generated rules. Some of the rules generated by J48 pruned tree model with all attributes are:

***RULE 1: IF*** *EDUCATION = NO AND TIME OF EXPOSURE <= 1377.333662 AND AGE <= 43 AND OUTMIGRATION = NO AND INMIGRATION = NO AND LITERACY = LI AND WINDOWS = NO AND OXEN = UK AND SEX = M AND ENVIRNMENT = H:* ***THEN*** *the class is Died (119.0/2.0).*

The first rule selected from the rules generated by J48 pruned tree model gives correct result of 119 out of 121 instances that it covers. From this, the likelihood of predictability of the individual to die or death attributed by the above predictors is about 98.4%.

***RULE 2: IF*** *EDUCATION = NO AND TIME OF EXPOSURE <= 1377.333662 AND AGE <= 43 AND OUTMIGRATION = NO AND INMIGRATION = NO AND LITER = LI AND WINDOWS = NO AND OXEN = UK AND SEX = M AND ENVIRONMENT= L:* ***THEN*** *the class is Died (38.0/1.0).*

The second rule selected from J48 tree pruned model is also shows that if the predictors in the above rule 2 are fulfilled, the chance of the person likely to die is 97.4%. All attributes are being constant in the first and second rules, the environment in which the adult lives matter the condition of adult health pattern and states that, living in rural highland and rural lowland is attributable to adult death. The domain experts accepted this rule by saying ‘rural lowlands and rural highlands are known socio demographic factors of adult death than urban’. When

comparing this rule with previous study in Butajira district says "most important socio-demographic factors that significantly associated with adult mortality are having no educated, the male sex, and living in the rural lowlands" [26].

Therefore, J48 pruned tree model of data mining technology reveals that if no education in family and the person is living in rural highland and lowland, the probability of experiencing adult death is 98.4% and 97.4% respectively with concomitant attributes in the above rules.

***RULE 3: IF*** *EDUCATION=COMPLETED PRIMARY:****THEN*** *the class is Alive(2855.0/33.0)*

***RULE 4: IF*** *EDUCATION = COMPLETED SECONDARY:* ***THEN*** *the class is Alive (1446.0/15.0)*

***RULE 5: IF*** *EDUCATION = FURTHER EDUCATION:* ***THEN*** *the class is Alive (153.0).*

Rule 3, 4 and 5 state that education alone matter the fate of the adult to live or to die without incorporating with other variables. As it has seen in the above rules, the probability of adult to survive is increases as the education attainment of the adult increases. Results from J48 pruned tree model of data mining algorithm reveals that the likely chance of adult to survive in completed primary school, completed secondary school, and further education is (98.9%, 99%, 100%) respectively. This in short can be explained as, the more the education the more the chance not to die in unfinished old age. The domain experts accepted this rule as education is a central role and mortality can be explained by education and adult with no formal education is more likely to die than adult with education.

Again comparing these rules with domain knowledge, study in rural Italy reported that men with college education were found to have significantly higher survival rates as compared to men who have no formal education [3]. According to Mitike et al. [26], education is one of the most important socio-demographic factors that affect adult health. Therefore, rule 3,4,and 5 are also strong(98.9%, 99%, 100%) respectively to predict adult mortality pattern.

***RULE 6: IF*** *EDUCA = UK AND OUTMIGRATION = NO AND INMIGRATION = YES AND LITER = LI AND RESIDENCE = Butajira area 04:* ***THEN*** *the class Alive (383.0/4.0).*

The rule gave a correct result for 383 of the 387 instances that it covers; thus its success fraction is 383/387(98.9%). This rule states that the likelihood of adult to live is 99% with concomitant

attributes in rule 6. In the domain area, this rule is a bit harsh to reality i.e. education unknown=>alive. However, assuming all the values of the other attributes, this rule is accepted as 'the residence in which the adult live determines adult's health condition. Previous study in Butajira District reveals that "living in Butajira town had a considerable survival advantage compared with the surrounding villages" [61]. Another study also reveals that "young adults from the rural highlands and lowlands had a higher risk of death than young urban adults" [45].

To this end, from the above rules 1 to 6, the J48 pruned tree algorithm with all attributes suggested that educations plays a considerable role as a root cause of adult death. Adult mortality is also associated with time of exposure, age, outmigration, inmigration, literacy, windows, oxen, environment and sex with the combination of education.

Thus, unfinished adult death by family or society due to early death of adult members can be minimized in formulating awareness creation for the people living at the rural area of the socioeconomic strata of the society through giving due emphasis for educational attainment in the area.

## 5.5 Error Rate of the Selected Model

Though the predictive performance of the selected model is promising 97.2% of accuracy for adult mortality prediction, the model commits 2.8% of the cases to classify wrongly to some other class.

The learning algorithm made bias to the majority class (alive) in our case such that in all the modes the predictive performance in identifying True Positive or alive cases of model is higher than identifying True Negative or die cases. This is because there is imbalanced between the two classes in the dataset. Consequently, the model tends to misclassify instances to some other class.

The other reason for misclassification is due to the fact that adult mortality status (alive or die) is based on the values of other attributes i.e. taking the similarity of the other attributes as a predominant predictive values. Table 5.6 presents instances of predicted and actual class.

**Table 5.6 Sample of Instances that Shows Predicted and the Actual Class**

| **Sample instances** | | **#1** | **#2** | **#3** | **#4** |
|---|---|---|---|---|---|
| **Attributes** | RESIDENCE | Meskan | Meskan | Butajira area 04 | Mjarda |
| | ENVIR | H | H | U | L |
| | REL | HE | CH | CH | HE |
| | SEX | M | M | F | F |
| | MARITAL | NM | TY | TY | PO |
| | TIMEX | 301 | 1411 | 686 | 222 |
| | LITER | LI | TY | TY | LI |
| | EDUCATION | NO | TY | NO | UK |
| | SOURCEW | PI | WU | PI | RI |
| | ROOF | CO | TH | CO | UK |
| | WINDOWS | YE | YE | YE | NO |
| | HOUSEOWN | OW | OW | OW | OW |
| | OXEN | NO | SI | UK | NO |
| | DISTHOSP | 2.9 | 2.9 | 0.8 | 17.4 |
| | AGE | 31 | 16 | 29 | 34 |
| | RADIUS | 3.41 | 3 | 3.41 | 3.41 |
| | TIMAD | 0 | 2 | 2.95 | 2 |
| | ROOMS | 2 | 1 | 2 | 1 |
| | INMIGRATION | YES | NO | NO | NO |
| | OUTMIGRATION | NO | NO | NO | NO |
| **Status** | **ACTUAL** | **Alive** | **Alive** | **Died** | **Died** |
| | **PREDICTED** | **Alive** | **Died** | **Alive** | **Died** |

As shown above Table 5.6, sample 1 and sample 4 were classified correctly as alive and died while actually they are alive and died respectively. Including sample 1 and sample 2, there are about 61095 instances that were correctly classified by the model. Sample 2 and sample 3 were misclassified as died and alive while actually they are alive and died respectively. Including sample 2 and sample 3, there are about 1774 instances that were misclassified to other class.

# CHAPTER SIX

# CONCLUSION AND RECOMMENDATION

## 6.1 Summary and Conclusion

Getting accurate data about patients' health care seeking conditions remained as a problem for a long time. In the times in which now we are living, application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as retail trade, health care, telecommunications, and banking. In particular, in the health care sector, data mining technology has been applied for patient survival analysis, prediction of prognosis and diagnosis, for outcomes measurement, to improve patient care and decision-making etc. Thus, the invention of ICT related technology (data mining), has made domain experts' tasks easy with regard to patients' health seeking data collection, processing, retrieval and application as well as predicting their future fate.

Although various research have been conducted in the area using traditional statistical tools in different aspects, data mining technology has not yet been elucidated in adult mortality condition in the area. This study attempted to explore data mining technology on adult predictive modeling using data base of BRHP that can help health care providers in the district to identify adults who are at risk for certain factors. Furthermore, such a predictive model can then be applied in assisting adult health care prevention and control activities in the region.

The hybrid, iterative methodology, was employed in this study which consists of six basic steps such as problem domain understanding, data understanding, data preparation, data mining, and evaluation of the discovered knowledge and use of discovered knowledge.

In order to generate interesting rule from the huge and massive data collected in the BRHP, a total of 43,864 instances in age group 15-60 years were taken using stratified simple random technique from each peasant association in respective of both classes(died and alive). Knowledge discovery in dataset was employed after having SMOTE technique has been done which is automatic operation by filter where minority classes are over sampled to make the target

attribute balance. In this particular research, the independent variables used are residence, environment, relation, sex, marital status, time of exposure, literacy, education, source of water, type of roof, windows in house, house own, oxen, distance from hospital, age, radius, timad, rooms, in migration, and outmigration with status attribute of the outcome variable.

The findings clearly suggest that these attributes have strong positive relationship with adult mortality in the program. All the selected attributes were used in the analysis using both decision tree and Naïve Bayes algorithms. Several models were built during experimentation that can predict the risk of adult mortality. Among the models, J48 pruned tree model with all attributes shows an interesting predictive accuracy result of 97.2% and 98.5% correctly predictive performance of individual as alive cases indeed they are alive. The best performing decision tree model was then chosen and evaluated using previously unseen records of adult.

The encouraging results obtained in this study shows that data mining is really a technology that should be considered to support adult health care prevention and control activities at the district of Butajira in particular, and at a national level in general.

In summary, it is concluded that the experiments presented in this study show that mortality can be reduced substantially by intervening in certain socio-economic and demographic effects so that probability of adult loose can be minimized. In formulating health policies, the people living at the rural of the socioeconomic strata of the society should get more importance in utilizing the education facilities to reduce avoidable mortality. Other differentiated predictors also need emphasis to come up adult mortality in the district. Enhancing data mining technology, particularly, the decision tree technique is well applicable to predict adult mortality patterns. Attributes which are recommended in the domain area like behavioral life style (smoking, chewing tobacco, and drinking alcohol), exposure to mass media, and type of fuel used for cooking were need to be integrated with existing variables in BRHP substantially to describe adult mortality in a country level.

Hence, based on the findings of this study, the following recommendations can be forwarded.

## 6.2 Recommendations

In this research work, efforts have been made to apply data mining technology to predict adult mortality patterns based on demographic, socioeconomic environmental and epidemiological factors.

Thus, based on the result of the research, the following recommendations are made the researcher would like to make the following recommendations. This will more enhance applicability of data mining technology in adult health prevention and control activities in inline with advocacy efforts of adult mortality reduction policy in rural communities of the country.
Thus, the following recommendations could be made considering as they are important issues for further research directions in adult mortality reduction strategies.

- The present study has considered epidemiological dataset to apply data mining in adult mortality prediction. Clinical data that have been gathered from different health care institutions should pay attention in adult mortality reduction. So that future study needs to discover knowledge and patters in clinical datasets and compare it with the result obtained using epidemiological datasets (BRHP).
- Although both the decision tree and Naives Bayes approaches resulted in an encouraging output, still performance improvement is expected. Hence, other classification algorithms such as neural networks and Bayesian network (Belief network) which have also been proved to be important techniques in the health care sector should be tested in order to investigate their applicability to the problem domain in the program by using the entire dataset
- In this research attempts were made to explore data mining technology to build adult mortality predictive modeling based on predefined classes (died, alive). It is appropriate to predict the survival years of the individual in the area corresponding to sample data available through data mining technology.
- Although both decision tree and Naïve Bayes reported promising results and hence could be applied in the area of adult mortality predictive modeling, decision tree tends to perform better. Thus, it would be more optimal for the Butajira Rural Health Program to employ the model developed with this technique.

- Results found from this research should be given attention so as to have a better decision making in the Butajira Rural Health Program particularly the program should give special attention to best attribute selected as mortality predictors such as education, time of exposure residence, outmigration, inmigration, and literacy.
- The possibility of incorporating the findings of this study with knowledge based system should be explored so that experts can consult the system in their problem solving and decision making process.

## REFERENCES


[1] Saikia N and Ram F. Determinants of Adult Mortality in India. Asian Population Studies. 2010; 6(2):153-171.

[2] Beaglehole R, Bonita R, Kjellstrom T. Basic Epidemiology: Switzerland. Geneva; 1993.

[3] Mesganaw F. Mortality and Survival from Childhood to Old Age in Rural Ethiopia. Umeå University Medical Dissertations, Sweden, Umeå University, SE-901 87 Umeå; 2008.

[4] Yamauchi F, Buthelezi T, Velia M. Impact of Prime Age Adult Mortality on Labor Supply: Evidence form Adolescent and Women in South Africa, USA: Washington DC: International Food Policy Research Institute, IFPRI; 2008.

[5] Yemane B, Stig W, Derege K, Anders E, Fikre E, Peter B, Lulu M, Tobias A, Negussie D, Ulf H, Yegomawork G, Atalay A, Kjerstin D. Establishing an Epidemiological Field Laboratory in Rural Areas Potentials for Public Health Research and Intervention. The Butajira Rural Health Program 1987-99. Ethiop J Health Dev.1999; 13:1-47.

[6] Yemane B and Peter B. Butajira DSS Ethiopia, Department of Community Health, Faculty of Medicine, Addis Ababa University and Department of Public Health and Clinical Medicine Umeå University**,** INDEPTH Monograph: Volume 1 Part C.

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

[8] Ngom P, Binka FN, Phillips JF, Pence B, Macleod B. Demographic Surveillance and Health Equity in Sub-Saharan Africa. Health Policy Plan. 2001; Dec 16(4):337-4.

[9] Kifle W, Yigzaw K, Kidist L. Epidemilogy: Lecture Note Series for Health Science Students. Jimma University: EPHTI, Ethiopia; 2003.

[10] Ruben D and Canlas Jr. Data Mining in Healthcare: Current Application and Issues. Thesis, Australia: Carnegie Mellon University; 2009.

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

[12] Mehamed Kantardzic J.B. Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc; 2003.

[13] Hand D, Mannila H, Smyth, P. Principles of Data Mining: UK, London: MIT Press; 2001.

[14] Amanuel D. Application of Data Mining Techniques to Predict Household Health Seeking Patterns: The Case of BRHP. MSc. Thesis. Addis AbabaUniversity, Addis Ababa; Ethiopia; 2011.

[15] Taddesse B. Mining Vital Statistics: The Case of Butajira Rural Health Program (BRHP). MSc. Thesis. Addis Ababa University, Addis Ababa, Ethiopia; 2011.

[16] Helen. T. Application of Data Mining Technology to Identify Significant Patterns In Census or Survey Data: The Case of Child Labor Survey in Ethiopia, MSc. Thesis. Addis Ababa University, Addis Ababa, Ethiopia; 2003.

[17] World Health Organization. World Health Statistics: Geneva, Switzerland WHO Press, Avenue Appia; 2010.

[18] Ngom P and Clark S. Adult Mortality in the Era of HIV/AIDS: Sub-Saharan Africa: Training Workshop On HIV/AIDS and Adult Mortality in Developing Countries. New York: Kenneth Hill; 2003.

[19] Sameh S. Sosena K, Petros Olango D, Banafsheh S. Improving Health Service Delivery: Ethiopia. The International Bank for Reconstruction and Development; 2009.

[20] Federal Democratic Republic of Ethiopia. Ministry of Health. National Adolescent and Youth Reproductive Health Strategy 2007-2015. Addis Ababa: Ministry Of Health; 2006.

[21] Federal Democratic Republic of Ethiopia. Health Sector Development Program IV, Ministry of Health. Addis Ababa: Ethiopia; 2011.

[22] MOFED(Ministry of Finance and Economic Development) of the Federal Democratic Republic of Ethiopia and the United Nations Country Team. Millennium Development Goals Report: Challenges and Prospects for Ethiopia. Addis Ababa: March; 2004.

[23] Patel I, Chang J, Srivastava J, Balkrishnan R. Mortality in The Developing World Can Pharmacists Intervene? Indian Journal of Pharmacy Practice. 2011, 4 (2):2-4.

[24] Yemane B, Stig W, Mesganaw F, Anders E, Wubegzier M, Ulf H, Alemayehu W, Fikru T, Mitike M, Negussie D, Abera K, Damen H, Fikre E, Peter B. A Rural Ethiopian Population Undergoing Epidemiological Transition Over a Generation: Butajira from 1987-2004. Scan J Public Health. 2008; 36:436-41.

[25] Duthe G, Pison G. Adult mortality in a rural area of Senegal: Non-communicable diseases have a large impact in Mlomp. GERMANY: Max Planck Institute for Demographic Research; 2008.

[26] Mitike M, Peter B, Yemane B, Bernt L. Mortality Decreases among Young Adults In Southern Central Ethiopia. Ethiop.J.Health *Dev.* 2008; 22 (3): 218-225.

[27] Cios Krzysztof J., Pedrycz Wiltod., Swiniarski Roman W. Kurgan Lukasz A. Data Mining: A knowledge Discovery approach. New York: Springer-Verlag Science Business Media LLC; 2007.

[28] Mehamed Kantardzic J.B. Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc; 2003.

[29] Milley A. Healthcare and Data Mining: Health Management Technology; 2000. 21(8), 44-47.

[30] Taft M, Krishnan R, Hornick M, Muhkin D, Tang G, Thomas S, Stengard P. Oracle Data Mining Concepts. USA: JD Edwards, PeopleSoft, and Retek Publication; 2005.

[31] Webb Geoffery I. Association Rules. In:Yen Nong,Editor. The handbook of Data Mining. New Jersey. USA: Lawrences Eralaum Associates, Inc; 2003.

[32] Ana Azevedo, Manuel Filipe S. Data Mining Standards, Knowledge Discovery in Databases: Data Mining; 2008.

[33] Rajaratnam Knoll J, Marcus R J, Levin-Rector A, Chalupka N A, Wang H, Dwyer L, Costa M, Lopez D A, Murray,LJC. Worldwide mortality in men and women aged 15–59 years from 1970 to 2010: A systematic analysis: Institute for Health Metrics and Evaluation, University of Washington, USA: Lancet. 2010; 30 (375):1704–20.

[34] Getu D and Fasil T. Biostatistics: Lecture Notes Series for Health Science Students. University of Gondar, Ethiopia: EPHTI; 2005.

[35] Bramer Max. Principles of Data Mining. Londan. Springer-Veralg Limited; 2007.

[36] Berry Michael W, Browne K M. Lecture Notes in Data Mining. USA: World Scientific Publishing Co. Pte. Ltd Inc. Rosewood Drive, Danvers, MA; 2006.

[37] Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. Third Edition. USA: Two crows Corporation; 2005.

[38] Berry J.A.M, Linoff G. Data Mining Techniques for Marketing Sales and Customer Support. New York: Wiley; 1997.

[39] Charniak E. Bayesian Networks without Tears. Publication of the American Association for Artificial Intelligence. California. USA: Lawrence Erlbaum; 1987.

[40] Witten H I. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco. Morgan Kaufmann, Elsevier Inc; 2005.

[41] Anthony S,Fauci H.(ed) On Harrison's Principles of Internal Medicine.Nwe York:McG;1997.

[42]. World Health Organization. Revised Global Burden of Disease (GBD) Estimates. Geneva: WHO; 2002.

[43] World Health Organization. The Ten Leading Causes of Death by Broad Income Group. Fact Sheet. World Health Organization; 2007.

[44] Murray. L. J C and Lopez D A. Global and Regional Cause of Death Patterns: Bulletin of the World Health Organization, 1994, 72 (3): 447-48.

[45]. Lulu K, Berhane Y, Tesfaye F. Sociodemographic Differentials of Adult Death in a Rural Population. Ethiop Med J; 2002. 40(4):375-85.

[46] Lulu K, Yemane B. The use of simplified verbal autopsy identifying causes of adult death in a predominantly rural population in Ethiopia. BMC Public Health. 2005; 5 (1):58.

[47] Bouckaert R, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, Scuse D. Weka Manual for Version 3-6-0. New Zealand: University of Waikato, Hamilton; 2008.

[48] Weiss Sholom M., Zhang Tong. Performance Analysis and Evaluation. In:Ye Nong, Editor. The Hand Book of Data Mining. New Jeresy. USA: Lawerence Erlbaum Associates Inc; 2003.

[49] Ifeachor C E, Hamadicharef B. Receiver Operating Curve Analysis in The Evaluation of Intelligent Medical Systems. UK: University of Plymouth. Drake Circus Plymouth PL4 8AA, Devon; 2004.

[50] Melanie C. Page, Sanford L Braver, David P. MacKinnon. Levine's Guide to SPSS for Analysis of Variance. London: Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey; 2003.

[51] Næss, Øyvind, Claussen, Bjørgulf and Smith, George D. Housing Conditions in Childhood and Cause-Specific Adult Mortality: The effect of Sanitary Conditions and Economic Deprivation on 55,761 men in Oslo, Scandinavian Journal of Public Health 2007, 1–7.

[52] Federal Ministry of Health: Essential Health Service Package. Federal Ministry of Health: August; 2005.

[53] Federal Ministry of Health. Main Module-11: Sample Handout on Health management for Environmental Health Technique. Adama, Ethiopia; August 2007.

[54] Federal Democratic Republic of Ethiopia: Ministry of Health, Health Sector Development Program IV, 2010/11 – 2014/15. First draft. Policy, planning and Finance General Directorate Addis Ababa, Ethiopia; 2010.

[55] Federal Democratic Republic Of Ethiopia: Combined Report (Initial And Four Periodic Reports). The African Commission on Human and Peoples' Rights Implementation of the African Charter on Human and Peoples' Rights. Addis Baba, Ethiopia; 2000.

[56] Habte D.Population and Health in Developing Countries.Washington DC: Published By the International Development Research Centre (1) World Bank; 2001.

[57] Foster L, Barkus J E, Yavorsky C. Understanding and Using Advanced Statistics. California. SAGE Publications Inc; 2006

[58] Chap T. Le. Introductory Biostatistics. Canada: John Wiley & Sons, Inc. Hoboken, New Jersey; 2003.

[59] Nitesh V, Chawla, Kevin W. Bowyer, Lawrence, O.Hall, W. and Philip K.SMOTE: Synthetic Minority Over Sampling Technique. Department of Computer Science and Engineering, ENB 188. University of South Floida; 2002.

[60] Larose Daniel T. Discovering Knowledge in Data-An Introduction to Data Mining. New Jersey USA: John Wiley & Sons Inc; 2005.

[61] Peter B, Mesganaw F, Wubegzier M, Anders E , Yemane B . From Birth to Adulthood in Rural Ethiopia: the Butajira Birth Cohort of 1987. Paediatr Perinat Epidemiol. 2008; 22(6):569-74.

# ANNEXES

## Annex 1: Calculation for Outlier Detection

Therefore, using quartile statistics from the above Table outliers can be detected as

**Rooms**= Lower adjusment =Q1-(1.5*IQR), Q1=1, Q3=2,IQR=1(2-1),

1-(1.5*1)=-5

=Upper adjusment = Q3+(1.5*IQR),Q3=2,1QR=1, 2+1.5=3.5

Therefore, all the values in Timad below -5 and above 3.5 is considered as outliers.

**Timad**= Lowere adjusmtent = Q1-(1.5*IQR), Q1=2,Q3=3,IQR=3-2=1,

2-(1.5*1)= 0.5

=Upper adjusment = Q3+(1.5*IQR),3+(1.5*1)= 4.5

Therefore, all values in below 0.5 and above 4.5 are considered as outliers.

**Radius** = Lowere adjusmtent = Q1-(1.5*IQR),Q1=3,Q3=4,IQR=1(4-3)

3-(1.5*1) = 1.5

=Upper adjusment = Q3+(1.5*IQR),4+(1.5*1)= 5.5

Therefore, all the values below 1.5 and above 5.5 are considered as outliers.

**Dishop** = Lowere adjusmtent = Q1-(1.5*IQR)= Q1=1.4, Q3=11.1, IQR=9.7

= 1.4-(1.5*9.7)=

-12.18

=Upper adjusment = Q3+(1.5*IQR), 11.1+(1.5*9.7)=25.35

There fore, all the values above 25.35 cosnsideered as outliers.

## Annex 2: Outputs of the Classifiers in Experimentation

### J48 pruned Decision Tree with selected Attributes

```
Scheme:       Weka.classifiers.trees.J48 -C 0.25 -M 2
Attributes:   11
Test mode 10-fold cross-validation

Number of Leaves  :     1054
Size of the tree :      2043
Time taken to build model: 7.27 seconds
Correctly Classified Instances       58896               93.6805 %
Incorrectly Classified Instances      3973                6.3195 %
Total Number of Instances            62869
=== Detailed Accuracy By Class = = =
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.96      0.107      0.944     0.96      0.952              0.964          Alive
                 0.893     0.04       0.922     0.893     0.907              0.964          Died
Weighted Avg.    0.937     0.084      0.937     0.937     0.937      0.964
=== Confusion Matrix ===
    a     b   <-- classified as
 39503  1646 |     a = Alive
  2327 19393 |     b = Died
```

### J48 Unpruned Decision Tree with all Attributes

```
Scheme:       Weka.classifiers.trees.J48 -U -M 2
Instances:    62869
Attributes:   21
Test mode  = 10-fold-cross validation
J48 unpruned  Tree
Number of Leaves  :     3046
Size of the tree :      3984
Time taken to build model: 5.12 seconds
Correctly Classified Instances       60922               96.9031 %
Incorrectly Classified Instances      1947                3.0969 %
Total Number of Instances            62869
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.978     0.048      0.975     0.978     0.976      0.973    Alive
                 0.952     0.022      0.958     0.952     0.955      0.973    Died
Weighted Avg.    0.969     0.039      0.969     0.969     0.969      0.973
=== Confusion Matrix ===
    a     b   <-- classified as
 40251   898 |     a = Alive
  1049 20671 |     b = Died
```

### J48 Unpruned Decision Tree with Selected Attributes

```
Scheme:       Weka.classifiers.trees.J48 -U -M 2
Instances:    62869
Attributes:   11
Test mode:    10-fold cross-validation
J48 unpruned tree
Number of Leaves  :       1573
Size of the tree :        3012
Time taken to build model: 6.08 seconds
Correctly Classified Instances       58867               93.6344 %
Incorrectly Classified Instances      4002                6.3656 %
Total Number of Instances           62869
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.956    0.102     0.947     0.956     0.952      0.96     Alive
                 0.898    0.044     0.916     0.898     0.907      0.96     Died
Weighted Avg.    0.936    0.082     0.936     0.936     0.936      0.96
=== Confusion Matrix ===
    a     b   <-- classified as
 39358  1791 |     a = Alive
  2211 19509 |     b = Died
```

### Naïve Bayes with All Attributes

```
Scheme:       Weka.classifiers.bayes.NaiveBayes –D
Instances:    62869
Attributes:   21
Test mode:    10-fold cross-validation
Naive Bayes Classifier
Time taken to build model: 2.96 seconds
Correctly Classified Instances       60187               95.734  %
Incorrectly Classified Instances      2682                4.266  %
Total Number of Instances           62869
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.981    0.087     0.955     0.981     0.968      0.983    Alive
                 0.913    0.019     0.962     0.913     0.937      0.983    Died
Weighted Avg.    0.957    0.064     0.957     0.957     0.957      0.983
=== Confusion Matrix ===
    a     b   <-- classified as
 40362   787 |     a = Alive
  1895 19825 |     b = Died
```

**Naïve Bayes with Selected Attributes**

```
Scheme:       Weka.classifiers.bayes.NaiveBayes –D
Instances:    62869
Attributes:   11
Test mode:    10-fold cross-validation
Naive Bayes Classifier
Correctly Classified Instances       59510               94.6571 %
Incorrectly Classified Instances      3359                5.3429 %
Total Number of Instances           62869
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.971     0.1       0.949     0.971     0.96       0.976    Alive
                 0.9       0.029     0.942     0.9       0.921      0.976    Died
Weighted Avg.    0.947     0.075     0.946     0.947     0.946      0.976
=== Confusion Matrix ===
     a     b   <-- classified as
 39955  1194 |     a = Alive
  2165 19555 |     b = Died
```

## Annex 3 : Description of the Selected Attributes

| No | Field name | Descriptions | Corresponding values |
|---|---|---|---|
| 1 | PA | villages of peasants | Bati, Dobena, Hobe, Mjarda, Bido, Dirama, Mmeskan, Wrib, Yeteker Butajira area 04 |
| 2 | ENVIR | Environment of the house hold is located | U =urban, L = lowland, H = highland |
| 3 | REL | Relationship with the head of house hold | HE = head of household, SP = spouse of head, CH = child of head/spouse, GP = parent of head/spouse,RE = other relative, NR = non-relative,<br>UK = unknown |
| 4 | SEX | F = female, M = male | F = female, M = male |
| 5 | MARITAL | Marital status during episode | MO = monogamous marriage,PO = polygamous marriage, NM = never married, DI = divorced, SE = separated, WI = widowed, TY = too young, UK = unknown |
| 6 | TIMEX | Days of exposure during episode | days of exposure during episode |
| 7 | LITER | Literacy during episode | LI = literate, RE = reading only<br>IL = illiterate, TY = too young to be at school, UK = unknown |
| 8 | EDUCATION | Educational status during episode | NO = no formal education, PR = completed primary school, SE = completed secondary school, TE = further education, UK = unknown |
| 9 | SOURCEW | Source of water during episode | RI = river, WU = well or spring unprotected, WP = well or spring protected, PI = urban supply (piped)<br>LA = lake or pond, OT = other<br>UK = unknown |
| 10 | ROOF | Type of roof during episode | TH = thatched roof, CO = corrugated roof, UK = unknown |
| 11 | WINDOWS | Windows in the house | YE = yes, NO = none, UK = unknown |
| 12 | RADIUS | Radius of circular house in metres | radius of circular house in metres |
| 13 | ROOMS | Number of rooms in house | number of rooms in house |
| 14 | HOUSEOWN | House ownership | OW = owned, KE = kebele or government<br>RE = privately rented, OT = other<br>UK = unknown |
| 15 | OXEN | Number of oxen owned by family | NO = none, SI = single animal, TW = two or more, UK = unknown |
| 16 | TIMAD | Number of timad of land owned by family | number of timad of land owned by family |
| 17 | DISTHOSP | Distance to Butajira | distance to Butajira km |
| 18 | AGE | Age of the individual | Age of the individual |
| 19 | INMIGRATION | Moving inside the population | Yes=if inmigration is yes, No=if inmigration is no. |
| 20 | OUTMIGRATION | Moving outside the population | Yes=if outmigration is Yes, No=if outmigration is no. |
| 21 | STATUS | Probability to die or survive | Die or Alive |

**Annex 4: Partial Decision Tree Generated for BRHP**

J48 pruned tree

---

EDUCA = UK

| OUTMIGRATION = NO

| | INMIGRATION = YES

| | | LITER = IL: Alive (595.0)

| | | LITER = LI

| | | | RESIDENCE = Meskan

| | | | | DISHOP <= 6.8: Alive (62.0/1.0)

| | | | | DISHOP > 6.8: Died (4.0/1.0)

| | | | RESIDENCE = Bati: Alive (161.0)

| | | | RESIDENCE = Dobena: Alive (122.0/2.0)

| | | | RESIDENCE = Bido: Alive (101.0/2.0)

| | | | RESIDENCE = Dirama: Alive (68.0)

| | | | RESIDENCE = Yeteker

| | | | | TIMEX <= 1148

| | | | | | AGE <= 49.498791: Alive (55.0)

| | | | | | AGE > 49.498791

| | | | | | | ROOF = TH: Died (2.0)

| | | | | | | ROOF = UK: Alive (2.0)

| | | | | | | ROOF = CO: Alive (0.0)

| | | | | TIMEX > 1148

| | | | | | SOURCEW = RI: Died (20.0/1.0)

| | | | | | SOURCEW = WU: Alive (2.0)

| | | | | | SOURCEW = PI: Died (0.0)

| | | | | | SOURCEW = LA: Died (0.0)

| | | | | | SOURCEW = WP: Died (0.0)

| | | | | | SOURCEW = UK: Died (0.0)

| | | | | | SOURCEW = OT: Died (0.0)

| | | | RESIDENCE = Wrib: Alive (91.0/1.0)

| | | | RESIDENCE = Mjarda: Alive (94.0/1.0)

| | | | RESIDENCE = Hobe: Alive (119.0/1.0)

| | | | RESIDENCE = Butajira area 04: Alive (383.0/4.0)

| | | LITER = TY: Alive (397.0/12.0)

| | | LITER = UK

| | | | AGE <= 32.495526: Alive (108.0/3.0)

| | | | AGE > 32.495526

| | | | | OXEN = UK

| | | | | | REL = HE: Alive (8.0)

| | | | | | REL = SP

| | | | | | | TIMEX <= 2178: Alive (7.0)

| | | | | | | TIMEX > 2178: Died (2.0)

| | | | | | REL = CH: Died (47.0/8.0)

| | | | | | REL = UK: Died (44.0/6.0)

| | | | | | REL = RE: Alive (18.0/1.0)

| | | | | | REL = NR: Alive (1.0)

| | | | | | REL = GP: Died (1.0)

| | | | | OXEN = SI: Alive (5.0)

| | | | | OXEN = NO: Alive (15.0)

| | | | | OXEN = TW: Alive (1.0)

| | | LITER = RE: Alive (22.0)

| | INMIGRATION = NO

| | | WINDOWS = NO

| | | | TIMEX <= 1383.609628

| | | | | AGE <= 30

| | | | | | MARITAL = UK

| | | | | | | AGE <= 29.018119

| | | | | | | | LITER = IL: Alive (243.0/3.0)

| | | | | | | | LITER = LI

| | | | | | | | | TIMEX <= 1338.725978: Alive (143.0/24.0)

| | | | | | | | | TIMEX > 1338.725978: Died (11.0/3.0)

| | | | | | | | LITER = TY

| | | | | | | | | ROOF = TH: Died (17.0)

| | | | | | | | | ROOF = UK: Died (0.0)

| | | | | | | | | ROOF = CO: Alive (4.0)

| | | | | | | | LITER = UK: Alive (90.0/4.0)

| | | | | | | | LITER = RE: Alive (2.0)

| | | | | | | AGE > 29.018119

| | | | | | | | AGE <= 29.99971: Died (43.0)

| | | | | | | | AGE > 29.99971

| | | | | | | | | LITER = IL: Alive (25.0)

| | | | | | | | | LITER = LI: Alive (15.0)

| | | | | | | | | LITER = TY: Died (24.0/1.0)