ADDIS ABABA UNIVERSITY

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MSc in Health Informatics

PREDICTING THE PATTERN OF UNDER-FIVE MORTALITY IN ETHIOPIA USING DATA MINING TECHNOLOGY: THE CASE OF BUTAJIRA RURAL HEALTH PROGRAM

BY

BE'EMNETU TEKABE

June, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

MSc in Health Informatics

PREDICTING THE PATTERN OF UNDER-FIVE MORTALITY IN ETHIOPIA USING DATA MINING TECHNOLOGY: THE CASE OF BUTAJIRA RURAL HEALTH PROGRAM

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

BY

BE'EMNETU TEKABE

ADDIS ABABA UNIVERSITY

SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

MSc in Health Informatics

PREDICTING THE PATTERN OF UNDER-FIVE MORTALITY IN ETHIOPIA USING DATA MINING TECHNOLOGY: THE CASE OF BUTAJIRA RURAL HEALTH PROGRAM

BY

BE'EMNETU TEKABE

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	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 our approval as university advisors.

_____

Getachew Jemaneh (Ato)

_____

Worku Tefera (Ato)

ACKNOWLEDGEMENTS

Above all, I would like to glorify the almighty GOD and St. Virgin Marry for giving me the ability to be where I am. You have done so much for me, O Lord. No wonder I am glad! I sing for joy, Amen!

Secondly, I would like to a very much grateful thank to my advisors Ato Getachew Jemaneh and Ato Worku Tefera for their constructive comments and overall guidance. But special thanks go to my mother W/ro Lakech W/Gebriel, without whom this research would have not been a success. Lakech, your helpful personality will always be a role in my heart.

I would also like to thank Dr. Alemayehu Worku, the AAU BRHP data manager, for allowing me to carry out this research using the required data from the BRHP database.

I am very much grateful to my brother Ato Migbaru Tekabe and Biniyam Admasu; and my sister W/ro Emiyou Tekabe for their care and understanding during my study times. I am also grateful to W/rt Anchialem Getahun to assist me morally and materially whenever I needed.

At last, but by no means the least, I would like to thank my friends for the constant assistance and encouragement they rendered to me since the time of my admission to the postgraduate program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Introduction: The under-five deaths in Ethiopia represent 48% of all mortality. More than half of the under-five deaths occurred during the first year of life, and 53% of these before 2 months of age. Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases.

Objective: The main objective of this study is to explore the potential applicability of data mining to predict the determinants, levels and pattern of under-five mortality in Ethiopia, particularly for the Butajira rural health program sites. This can greatly support for policy makers, planners, and healthcare providers working on the control of under-five children mortality in Ethiopia.

Methods and Material: The methodology used for this research was a hybrid six-step Cios Knowledge Discovery Process. The required data was collected from Butajira rural health program database covering the period 1987-2008. The researcher used two popular data mining algorithms (C4.5 J48 Decision Trees and Naïve Bayes Classifier) to develop the predictive model using a larger dataset (11,600 cases). The researcher also used a 10-fold cross validation and 90% split test mode for data mining methods of the two predictive models for performance comparison purposes.

Results: The results indicated that the decision tree (J48 algorithm) is the best predictor with pruned parameter of the tree of 90% split test mode; it has 97.49% accuracy on the holdout dataset (this predictive accuracy is better than any reported in the literature), Naïve Bayes Classifier came out to be the second with supervised discretization has 96.67% accuracy.

Conclusion: The results from this study were very capable and confirmed the belief that applying data mining techniques could indeed support a predictive model building task that predicts the pattern of under-five mortality in Ethiopia; particularly for Butajira rural health program sites are possible. In the future, more classification studies by using a possible large amount of Butajira rural health program demographic and surveillance sites dataset records with epidemiological information and employing other classification algorithms, tools and techniques could yield better results.

ACRONYMS

AAU: Addis Ababa University

AI: Artificial Intelligence

ARFF: Attribute Relation File Format

AUC: Area under the ROC

BRHP: Butajira Rural Health Programme

CRISP-DM: CRoss-Industry Standard Process for Data Mining

CSV: Comma Separated Value

DBase: Database

DM: Data Mining

DSA: Demographic and Surveillance Area

DSS: Demographic and Surveillance Sites

EDHS: Ethiopian Demographic and Health Survey

GPS: Global Positioning System

HIV: Human Immunodeficiency Virus

HSEP: Health Service Extension Program

ID3: Interactive Dichotomizer 3

INDEPTH: The International Network for the continuous Demographic Evaluation of Populations and Their Health in developing countries

KD: Knowledge Discovery

KDD: Knowledge Discovery in Databases

KDP: Knowledge Discovery Process

KM: Knowledge Mining

MDG: Millennium Development Goal

NGO: Non Governmental Organization

NSCSE: National Strategy for Child Survival in Ethiopia

OLAP: On-Line Analytical Processing

ROC: Receiver Operating Characteristics

SEMMA: Sample, Explore, Modify, Model, and Assess

SMOTE: Synthetic Minority Oversampling Technique

SNNPRS: Southern Nations, Nationalities and Peoples Regional State

SQL: Structured Query Language

UNDP: United Nation Development Program

VA: Verbal Autopsies

VUS: Volume under the Surface

WEKA: Waikato Environment for Knowledge Analysis

WHO: World Health Organization

CHAPTER ONE

INTRODUCTION

1.1 Background

In Ethiopia around 472,000 children die each year before dying between birth and their fifth birthdays. There are awful reports that piece of evidence places Ethiopia sixth amongst the countries of the world in terms of the absolute number of child deaths. Yet, there are effective and proven tools which can be used to achieve the Millennium Development Goal (henceforth MDG) of reducing child deaths by two-thirds by 2015, taking 1990 as a benchmark [1].

The MDG-4 United Nation's (henceforth UN) greatest achievements, arguing that global goals and benchmarks have influenced policies and outcomes in many countries to reduce under-fives mortality by two-thirds between 1990 and 2015 has come into centre of attention in recent years as a galvanising force to align global and national efforts towards poverty reduction and better health sector promotion [2].

Information obtained from Ethiopian 2011 Demographic and Health Survey (henceforth DHS) shows that a rapid decrease in infant and under-five mortality during the five years prior to the survey compared to the period 5-9 years earlier. The levels are also considerably lower than those reported in the 2005 EDHS. However, the under-fives mortality rates are also highest among the world countries. For example, infant mortality has decreased by 23 percent, from 77 to 59 deaths per 1,000 births, while under-five mortality has decreased by 28 percent, from 123 to 88 per 1,000 births [3].

According to the Central Statistical Authority (henceforth CSA) [4], around 90% of under-five mortality in Ethiopian children is caused by pneumonia, neonatal causes (prematurity, asphyxia and neonatal sepsis), malaria, diarrhoea, and measles. Malnutrition is the underlying cause of death in about 57% of these deaths, and 11% are associated with HIV infection. The levels of under-five mortality are also worsened particularly by poverty, inadequate maternal education, lack of potable water and sanitation, high fertility and inadequate birth spacing [1].

Recent information shows that the occurrence rate of under-five mortality is still very high. For the five years immediately preceding the EDHS (corresponding roughly to 2006–2010); the infant mortality rate was 59 deaths per 1,000 live births. The estimate of child mortality is 31 deaths per1000 children surviving from 12-59 months of age, while the overall under-five mortality rate for the same period is 88 deaths per 1,000 live births. Sixty-seven percent of all deaths to children under-five in Ethiopia take place before a child's first birthday [3]. These figures are very high even compared to other developing countries which are severely affected by the problem of the under-five mortality.

The government of Ethiopia has given high priority child survival interventions. This decision has been taken in a context which strongly supports such action.

Although community-based health information is vital for rational health planning, evaluation, and intervention, one of the fundamental challenges of many developing countries in health care delivery is lack of adequate health information system which collects, compiles, analyses, interprets, and disseminates health related information for planning and decision making [1].

The Butajira Rural Health Programme (henceforth BRHP) was initiated in mid 1986 with a complete census of the ten sampled kebeles (kebele is the smallest administrative unit in Ethiopia) in the former district of Meskan and Mareko. Soon after, by January 1987, a Demographic and Surveillance Sites (henceforth DSS) with continuous registration of vital events was initiated. The major aims were to develop and evaluate a system for continuous registration of births and deaths, to generate valid data on fertility and mortality and to provide a study-base for essential health research and intervention in the area [5].

The BRHP is primarily a collaborative research undertaking between the Department of Community Health, Faculty of Medicine, Addis Ababa University (henceforth AAU) in Ethiopia, and the Division of Epidemiology, department of Public Health and Clinical Medicine, Umeå university, Sweden. And also, the collaboration started as an individual doctoral study project [5].

During the first ten years of monitoring at BRHP DSS site, a total of 5,143 deaths and 15,667 births were registered in the area, from a total of 336,074 person-years of follow-up. Thus, by

relating the observed total number of deaths to this study base, the crude mortality rate was 15.3 per 1,000 person-years.

Moreover, under-five mortality in Ethiopia exhibits a considerable variation between rural and urban areas. To be specific, under-five deaths are overrepresented in rural parts of the country than their urban counter parts. Thus, among live births 4.2% are estimated to die during the first two months of life, 8.0% before one year, 16.6% before 5 years, 36% before 15 years, and 56% before 65 years. There were substantial variations between areas with regard to under-five mortality, with rates ranging from 80 per 1,000 person-years in the urban area to 219 per 1,000 person-years in the rural area [5].

Data Mining (henceforth DM) can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. The process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns. Lastly, some of the DM algorithms make use of rules, which are required for categorization. Rules are obtained based on patterns present in the training data set, which are extracted by the various DM algorithms [6].

Medical informatics plays a very important role in the use of clinical data as well as epidemiological data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place.

According to Han and Kamber [7], the major reason that “DM has attracted a great deal of attention in the information industry as well as academic area in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge”. DM tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research [7].

As a particular application of predictive modeling, this area raises important considerations about how future health determinants will change e.g. evolving ecologies, shifting patterns of disease, new health challenges, new vectors of disease, new socio-cultural determinants of ill health, and thermal stress from poorly insulated houses. As the researchers had seen how DM applications

could be used in early detection of diseases, prevention of deaths, the improvement of diagnoses and even detecting fraudulent health claims [8]. However, there are caveats to the use of DM in healthcare. For reasons of familiarity and availability of electronic data, the researcher chose BRHP to conduct the research study.

In light of this, therefore, predicting the pattern of under-five mortality in Ethiopia using epidemiological DSS data particularly socio-demographic and public health determinants (risk factors) that are associated with under-fives deaths on survival status of Ethiopian children can assist the effort towards alleviating the burden of the under-five mortality. Thus, this paper focuses on predicting the pattern of under-five mortality in Ethiopia, specifically for BRHP DSS area.

1.2 Statement of the Problem

Much of the current burden of mortality in children younger than five years in low-income countries is preventable, if effective coverage of available cost-effective interventions can be achieved [9].

In the Ethiopian case, studies have shown that the under-five deaths represent 48% of all mortality. More than half of the under-five deaths occurred during the first year of life, and 53% of these before 2 months of age. From the age-specific mortality rates, we can estimate the cumulative mortality throughout life. Thus, among live births 4.2% are estimated to die during the first two months of life, 8.0% before one year, 16.6% before 5 years, 36% before 15 years, and 56% before 65 years [3].

Recent information shows that a large proportion of children have died from Malaria (7%), Pneumonia (18%), Diarrheal Diseases (11%) and Neonatal conditions (35%); each account for the major causes of under-five deaths in Ethiopia [9].

Moreover, understanding the current determinants of child mortality is essential to inform policies and strategies to accelerate the reduction of child mortality. It is often associated with poverty (the lowest quintile is associated with 32% more child mortality than the highest), maternal education, maternal fertility characteristics [the under-five mortality rate is significantly higher for mothers under the age of 20 (225 deaths per 1,000 compared to 179 for mothers in

their twenties)], maternal under-nutrition, intervals between births, access to adequate safe water and basic curative health services. The expansion of family planning programs across the country is expected to positively impact on some of the challenges associated with reducing under-five children mortality [9].

In Ethiopia, the practical challenge for policy makers, planners, and health care providers, working in primary health care disease prevention and control activities is lack of timely and reliable health information on the health status of defined population groups [10].

Data were handled using custom-made software based on the dBase system. Events registered by the BRHP are birth, death, marriage, new household, out-migration, in-migration, and internal move. Household and environmental variables are measured during the censuses [5].

The major aims of BRHP were to develop and evaluate a system for continuous registration of births and deaths, to generate valid data on fertility and mortality and to provide a study-base for essential health research and intervention in the area [11].

In this thesis, the researcher dealt with text (document with paper) data only. A few of those text data problems like automated classification was solved with the help of context based text classification. Typical approaches extract features out of the data that is submitted in the BRHP. Data entry for the DSS was firstly processed as text strings, but since 1994 has been performed, using software based on the dBase IV platform. This program as developed for Butajira includes procedures for automatic consistency checking as well as more sophisticated facilities for data management and retrieval. Since an indigenous calendar is used in Ethiopia, which runs 2,809 days behind the international equivalent and has 13 months in a year, there are serious obstacles to using proprietary packages for handling longitudinal data. Data entry was currently done in Butajira, which enables any inconsistent questionnaires to be sent back to the field immediately. This is a significant improvement over earlier periods when data operations were centralised in Addis Ababa; and also the reconciliation tasks for the project activities were done at AAU [5].

Data have been manipulated and analysed using dBase, Epi-Info and the Cohort program developed by Umeå University, which performs person-time based analyses of events in dynamic cohorts. National and international publications and scientific conferences have been

the main routes of dissemination of information. Community feedback meetings have been held periodically [5].

As with other aspects of DM, while technological capabilities are important, other implementation and oversight issues can influence the success of a project's outcome. The issues are data quality, interoperability, mission creep or the use of data, and privacy [12].

According to UN's MDG report, the improvement of under-five children's health is an essential component of the Health Sector Development Programme (henceforth HSDP) III that focuses on poverty related health conditions. HSDP III that ends in the middle of 2010 envisaged a reduction of the mortality rates of children under-five from 123/1,000 to 85/1,000 and the infant mortality rate from 77 to 45 per 1000 live births. This prognosis is based on an increased coverage of maternal, newborn and child health, nutrition and wash related interventions [9].

Achieving the MDG-4 for child survival in Ethiopia demands focused and coordinated action to improve nutrition, to strengthen health systems, and to reduce inequities in access to effective interventions against the diseases that kill under-five children [1].

Previous investigations were attempted to show in the areas of under-five children mortality have proved the applicability of DM as well as using simple statistical method. Shegaw [11] has proved the applicability of DM technology in predicting child mortality at BRHP. The researcher has proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities. Abera [13] has also tried to show the determinants of child mortality at BRHP by using retrospective cohort study design. Infant and under five mortality rates were 83.9 and 118 deaths per 1000 live births. Excess mortality was observed in female children than in males; moreover, multiple births were at increased risk of dying than singleton.

The problems of previous research efforts regarding to the under-five mortality were conducted not only related to the small proportion of dataset used, but also the data analysis was conducted by using simple statistical techniques (such as logistic regression and verification), applying simple algorithm, and /or lack of standard DM tools. The researcher believed there are at least 10 years differences of the data collected from the database. So, the researcher tried to demonstrate

different DM methods, techniques and tools as well as the gap identified, study design, model classifiers, automated pattern recognition, and attempts to uncover pattern in the dataset was presented on the following research investigation.

To this end, the study attempts to explore, investigate, and answer the following main research questions:

- What are the major determinant attributes that contributes to the pattern of the under-five mortality?
- Which DM tools, techniques and models are more appropriate to predict the pattern of the under-five mortality in Ethiopia?
- Is there any pattern that can be extracted from BRHP DSS data for the prediction of under-five mortality?

1.3 Objectives of the Research

1.3.1 General Objective

The general objective of this research is to explore and design a predictive model using DM technology to predict the determinants, levels and pattern of the under-five mortality in Ethiopia.

1.3.2 Specific Objectives

- To extract the dataset required for the problem domain and analysis from the database of BRHP DSS.
- To identify the major determinant attributes for the occurrences of under-five mortality.
- To design suitable DM method for creating predictive model using BRHP DSS dataset of the under-five mortality data.
- To build DM models on the pre-processed dataset.
- To measure the performance of the generated model using confusion matrix and ROC analysis.
- To evaluate different algorithms like, decision tree (J48) and Naïve Bayes classifier.
- To generate rules for the predicted model of the classified classes.
- To extract interesting patterns that is relevant to the problem domain.

1.4 Research Methodology

1.4.1 Research Design

The researcher used a hybrid Knowledge Discovery Process (KDP) model to achieve the goal of building predictive model using DM technique. This methodology was selected among different methodologies like KDD, CRISP-DM, SEMMA, 5 A's, etc. due to the benefits and the needs of the academic research community, providing a more general, research-oriented description of the steps [14]. All anticipated models also emphasize the iterative nature of the model, in terms of many feedback loops and repetitions, which are triggered by a revision process.

Based on the hybrid model of the Cios six steps methodology, the following procedures are identified in order to predict the pattern of under-five mortality in Ethiopia, particularly for the BRHP DSS area.

1.4.2 Understanding the Problem Domain

It is the essential goal of the DM research, which can support to understand the research area or problem domain. In addition, the significance of the impact of determining goals on the outcome of the DM experiment is highlighted. The BRHP's DSS goals were converted to DM goals.

In this stage the researcher has tried to review the relevant literature to assess DM technology, both concepts and techniques, and researches in this field. Various books, journals, magazines, and papers from the internet pertaining to the subject matter of DM and KDP were reviewed to understand the potential applicability of DM to predict the pattern of under-five mortality in Ethiopia.

1.4.3 Data Understanding

The major goal of this step is to understand the data sources, data parameters and quality of data. The potential source of data to be used to undertake this research was the BRHP DSS database. The pre-classified historical data was collected about under-five children by the BRHP. As a result, the researcher identified two main sources of data stored in electronic format. The first source of data to be identified the main database of the project. This is a database, which is updated every three months and incorporates all new information gathered about the study

population through census and the surveillance system. The second source of data to be identified was the separate surveillance data, which is a definitive version of the main database, but, it was purposefully cleaned and prepared to provide a separate dataset for those who are interested for the analysis of the population-based data gathered by the BRHP DSS area. Therefore, the researcher tried to collect the data, assure the data quality and reconcile the collected data based on the INDEPTH standards.

1.4.4 Data Preparation

This is one of the crucial steps to construct dataset used for modeling by Waikato Environment for Knowledge Analysis (henceforth WEKA) software. At this stage, all necessary tasks needed to perform DM are finalised. DM techniques, tools and algorithms were decided. The data sets are pre-processed for specific DM tasks. Pre-processing includes selecting, cleaning, deriving, integrating, and formatting data in order to apply specific DM tasks. There are a number of possible DM techniques such as classification, clustering, association rule, regression, and others. Therefore, the researcher selected decision tree and Naive Bayes algorithms, pre-processed the data as well as the following tools were also used in the data preprocessing phases: SPSS, MS-Office 2007, WEKA and Note Pad softwares.

1.4.5 Data Mining

To build a predictive model from the cleaned data, WEKA-DM software was used. At this stage, a DM model is developed to represent various characteristics of the underlying dataset. A number of iterations in fine-tuning the model were taken place. The researcher typically iterated through process to find a best-fitted DM model for the data by adjusting various parameters of the model (e.g. threshold values).

This model was used to describe the hope of discovering novel and useful patterns and trends in data; or used to predict future or unknown values .Therefore, J48 Decision Tree and Naïve Bayes algorithms were applied to predict the pattern of under-five mortality in Ethiopia.

1.4.6 Evaluation of the Discovered Knowledge

At the evaluation step of the DM experiment, the results of the particular DM tasks was visualised and interpreted. Although DM tasks reveal patterns and relationships, this by itself is not sufficient. Domain knowledge and DM expertise is required to interpret, validate and identify interesting and significant patterns. The DM team incorporates domain expertise and DM expertise in evaluating and visualising models in order to identify interesting patterns and trends.

In this phase, the investigator tried to evaluate the J48 decision tree and Naïve Bayes classifier model performance and evaluation by means of confusion matrix as well as ROC analysis and also discussion on the generated rules or models with domain experts from AAU and BRHP DSS area and literatures.

1.4.7 Use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is reported. Therefore, the researcher tried to disseminate the discovered knowledge by means of thesis defense presentation, availability of the thesis in the libraries and presentation in Ethiopian Public Health Association.

1.5 Scope and Limitation of the Research

The scope of this research is to train and test the potential applicability of DM technology in supporting under-fives survival strategies at the BRHP DSS area. For a good prediction or classification the learning algorithms must be provided with a good training set from the BRHP and DSS database which rules or patterns are extracted to help to classify the testing dataset. The output of this research can be used as an input for the Federal Ministry of Health (FMOH) under-fives survival strategies. The researcher initially intended to evaluate the potential applicability of DM technology in the Ethiopian healthcare sector at large. However, the DM investigation task was limited only to one site, BRHP, because of the availability of organized data.

In conducting this research, the researcher has faced lack of availability of literature on the application of DM for BRHP DSS data. And also the limitation of this study was lack of prior experience on DM techniques as well as to the problem domain knowledge was also the restrictions faced by the investigator.

1.6 Significance of the Study

The significance of this research is to predict the pattern of under-five mortality in Ethiopia. The findings of the study can be of used to the FMOH, professionals associations, and other stakeholders such as NGOs in designing and implementing child health intervention programs and projects**.** The main contribution of this thesis:

- Proportionally reduce the neonatal, infant and child mortality rates while achieving the under-five children survival strategies.
- Will ensure the greatest possible reduction of under-fives mortality among the children of the poorest and most marginalized sections of the population.
- Contribute to the reduction of under-fives mortality to achieve the MDG by 2015.
- Will ensure the availability of quality essential health care for children in the community and health facilities.
- Supports Primary Health Care (PHC) providers, planners, policy makers and decision makers for a better efficiency.
- Will also serve as a base for conducting further investigations in the area of the problem domain.

1.7 Dissemination of the Research

The final report of this research will be presented and submitted to School of Public Health and Information Science at AAU in partial fulfillment of the requirements for MSc degree in Health Informatics. The summary of the final report will be submitted to stakeholders who have a keen interest on the subject matter. The result will also be presented at annual conferences of Ethiopian Public Health Association and other professional associations.

1.8 Thesis Organization

This research report is organized into six chapters. The first chapter briefly discusses the background of the study, statements of the problem, general and specific objectives of the study, research methodology, scope and limitation, significance of the research, and application of the results of the research. Chapter two and three, review the DM technology and the methods for mining under-fives mortality DSS data respectively. The concepts pertaining to the DM technology and its application in the problem are reviewed in chapter two. Moreover, the previous research in which a researchable gap was left and become the concern of this research is reviewed in this chapter, chapter two. Chapter three is dedicated for the discussion of basic issues, tools, techniques and algorithms that can be relied on the under-fives mortality at the districts of BRHP (the problem area). Chapter four explains the Business Understanding and Data Preprocessing used in this research. Chapter five presents the experimentation and analysis phase of the study at hand. Results of the classification experiments were also discussed here. Finally, chapter six provides conclusion of the research, and also presents recommendation for future work.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Under-five Child Mortality

The review of literature for under-five children mortality showed that a number of attributes or variables affecting under-five mortality. However, the predictors of child mortality are changing through time. Since, the infrastructures facilities and awareness are changing day by day. Hence, it is necessary to give more emphasis on using the current data to identify the segment of residents where BRHP need to be strengthening in order to achieve the goal and objectives of the research project.

As mentioned earlier, approximately 11 million children under age of five die annually in the world as a whole, of which over 10 million are in the developing world. A large proportion of these deaths are preventable and uncounted. To this end, a realistic picture of an epidemiological profile and intervention developments require an understanding of the determinants and patterns of under-five mortality [21]. Moreover, a relatively less expensive and feasible method that can yield reliable and valid data is necessary. Even though, the fact that many studies that have been done to determine factors associated with child deaths, no sound methods, techniques and tools were used [13]. Like other studies from other parts of Africa, the studies on under-five mortality in Ethiopia have also tended to operate, almost invariably, within the traditional public health, socioeconomic and demographic framework by concentrating largely on such factors as birth intervals, birth order, traditional practices, educational level of the mother, residence, vaccination status and other socio-economic and demographic indicators.

Moreover, there are significant variations in mortality by socio-economic determinants. The under-five mortality rate for the poorest 20% of the Ethiopian population is 32% higher than that

for the richest (20%). Poverty not only affects food supply and access to health care, but it is linked to higher fertility rates, which in turn is associated with the twisting of poverty. The under-five mortality rate for children who live in rural areas is 30% higher than that for children who live in urban areas [1]. Most of the children that survived their first month of life did not celebrate their second birthday**.** Distance to the nearest modern health facility appears to be a good predictor of place of death. In general, as distance increases the number of deaths occurring at home also increases.

Children for whom the preceding birth interval was less than two years had a mortality rate of 272/1000, compared with a mortality rate of 96 for whom the interval was four or more years. While the effect may be reduced by controlling for other socio-economic determinants (e.g. education) and death of the preceding child, it is likely that birth intervals play an important role in determining childhood mortality in Ethiopia. In 2000, the median birth interval was 34 months, and 20% of all preceding birth intervals were less than two years [1].

About 28% of Ethiopian families have access to adequate and safe water, and 11.5% have access to excreta disposal. There is ample evidence that access to adequate and safe water and sanitation can influence child mortality and, therefore, these major determinants must be addressed in developing sustainable preventive interventions [1]. An innovative methodology, which has been recommended for use in other parts of the developing world because of its reliability in explaining under-five mortality, was tried to use in this research study.

2.2 Health Informatics

Health care is a very research intensive field and the largest consumer of public funds in developed countries. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no longer ignore these emerging tools. This resulted in uniting of healthcare and computing to form health informatics (Health informatics exists since the 1950's). This is expected to create more efficiency and effectiveness in the health care system, while at the same time, improve the quality of health care and lower cost.

Health informatics is an emerging field. It is especially important as it deals with collection, organization, storage of health related data. With the growing number of patient and health care

requirements, having an automated system will be better in organizing, retrieving and classifying of medical data. Physicians can input the patient data through electronic health forms and can run a decision support system on the data input to have an opinion about the patient's health and the care required. An example in the advances in health informatics can be the diagnosis of a patient is health by a doctor practicing in another part of the world. Thus, healthcare organizations can share information regarding a patient which will cut costs for communication and at the same time be more efficient in providing care to the patient [56].

There are other issues like data security and privacy, which is equally important when considering health related data. Thus, health informatics deals with "biomedical information, data, and knowledge with their storage, retrieval, and optimal use for solving problem and decision making process" [56]. This is a highly interdisciplinary subject where fields in medicine, engineering, statistics, computer science and many more come together to form a single field.

With the help of smart algorithms and machine intelligence we can provide the quality of healthcare by having, problem solving and decision-making systems. Information systems can help in supporting clinical care in addition to helping administrative tasks. Thus, the physicians will have more time to spend with the patients rather than filling up manual forms [56].

2.3 Data Mining and Knowledge Discovery in a Database

This thesis is concerned with DM: extracting useful insights from large and detailed collections of data. With the increased possibilities in modern society for companies and institutions to gather data cheaply and efficiently, this subject has become of increasing importance. This interest has inspired a rapidly maturing research field with developments both on a theoretical, as well as on a practical level with the availability of a range of commercial tools.

Traditional methods (methods used before computers were introduced into healthcare) use manual analysis to find patterns or extract knowledge from the database. For example in the case of health care, the health organizations (e.g. the center for disease control in the USA) analyze the trends in diseases and the occurrence rates. This helps health organizations take precautions in future in decision making and planning of health care management [35].

DM refers to extracting or mining “knowledge” from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, “DM" should have been more appropriately named “knowledge mining from data", which is unfortunately somewhat long. “Knowledge Mining (henceforth KM),” a shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a bright term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both “data” and “mining” became a popular choice. Many other terms carry a similar or slightly different meaning to DM, such as KM from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [35, 19].

On the other hand, as the writers adopt the convention that DM refers to the act of extracting patterns or models from data (be it automated or human-assisted). However, many steps precede the DM step: retrieving the data from a large warehouse (or some other source); selecting the appropriate subset to work with; deciding on the appropriate sampling strategy; cleaning the data and dealing with missing fields; and applying the appropriate transformation, dimensionality reduction, and projections. The DM step then fits models to, or extracts patterns from, the preprocessed data. However, to decide whether this extracted information does represent knowledge, one need to evaluate this information, perhaps visualize it, and finally consolidate it with existing (and possibly contradictory) knowledge. Obviously, these steps are all on the critical path from data to knowledge. Furthermore, any one-step can result in change in the preceding or succeeding steps, often requiring starting from scratch with new choices and settings [31].

By definition, Knowledge Discovery in Database (henceforth KDD) is an interdisciplinary field that brings together researchers and parishioners from a wide variety of fields. The major related fields include statistics, machine learning, artificial intelligence and reasoning with uncertainty, databases, knowledge acquisition, pattern recognition, information retrieval, visualization, intelligent agents for distributed and multimedia environments, digital libraries, and management information systems. The remainder of this section briefly outlines how some of these relate to the various parts of the KDD process. We focus on the main fields and hope to clarify to the reader, the role of each of the fields and how they fit together naturally when unified under the

goals and applications of the overall KDD process. A detailed or comprehensive coverage of how they relate to the KDD process would be too lengthy and not very useful because ultimately one can find relations to every step from each of the fields [30].

Moreover, the main advantage of the discovery process is that no hypotheses are needed and knowledge is extracted from the data without previous knowledge. KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular DM method. Knowledge Discovery (henceforth KD) and DM are powerful data analysis tools. DM is an important tool for the mission critical applications to minimize, filter, extract, or transform large databases or datasets into summarized information and exploring hidden patterns in KD [14, 35, and 19]. DM and KDD are defined as a non-trivial discovery process of valid, new, useful and accessible patterns [22].

Lastly, when as writers explains encounter patterns within a database the researchers state the findings (patterns or rules) as DM, information retrieval or knowledge extraction and so on. The term DM is used mostly by statisticians, data analysts and the management information systems (MIS) professionals [56]. The difference between DM and KD is that the latter is the application of different intelligent algorithms to extract patterns from the data whereas KD is the overall process that is involved in discovering knowledge from data. There are other steps such as data preprocessing, data selection, data cleaning, and data visualization, which are also a part of the KDD process. Many people treat DM as a synonym for another popularly used term, KD from Data, or KDD. Alternatively, others view DM as simply an essential step in the process of KD [7]. Hence, in the definition as the writer adopt, DM is just a step in the overall KDD process.

2.4 Data Mining and Data Warehouse

In predictive models, the values or classes as researchers are predicting are called the response, dependent or target variables. The values used to make the prediction are called the predictor or independent variables. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. On the other hand, descriptive techniques are sometimes referred to as unsupervised learning because there is no already known result to guide the algorithms [31].

The relevance of the field of databases to KDD is obvious from the name. Databases provide the necessary infrastructure to store, access, and manipulate the raw data. With parallel and distributed database management systems, they provide the essential layers to insulate the analysis for the extensive details of how the data is stored and retrieved. We focus here only on the aspects of database research relevant to the DM step [7, 31]. A strongly related term is on-line analytical processing (henceforth OLAP), which mainly concerns providing new ways of manipulating and analyzing data using multidimensional methods. This has been primarily driven by the need to overcome limitations posed by SQL and relational DBMS schemes for storing and accessing data [30].

Supporting operations from the DM perspective has an emerging research area in the database community. In the DM step itself, new approaches for functional dependency analysis and efficient methods for finding association rules directly from databases have emerged and are starting to appear as products. In addition, classical database techniques for query optimization and new object-oriented databases make the task of searching for patterns in databases much more reasonable [30].

Data warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involves data cleaning, data integration and data transformation and can be viewed as an important preprocessing step for DM. Hence, the data warehouse has become an increasingly important platform for data analysis and OLAP and will provide an effective platform for DM [24]. Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse systems are valuable tools in today's competitive, fast-evolving world. Many people feel that with competition mounting in every industry, data warehousing is the latest must-have marketing weapon-a way to retain customers by learning more about their needs [26, 30].

Moreover, data warehouses have been defined in many ways, making it difficult to formulate a rigorous definition. Loosely speaking, a data warehouse refers to a database that is maintained separately from an organization's operational databases. Data warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data for analysis [7].

According to Immon [26], a data warehouse is “an integrated collection of data about a collection of subjects (units), which is not volatile in time and can support decision taken by the management”. The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

In sum, a data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. A data warehouse is also often viewed as architecture, constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making [7].

2.5 Data Mining, Artificial Intelligence and Statistics

DM takes advantage of advances in the fields of Artificial Intelligence (henceforth AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification.

AI techniques for reasoning, especially techniques form the uncertainty in AI community and graphical models for Bayesian modeling and reasoning provide a powerful alternative to classical density estimation in statistics [30]. These techniques have the advantage of allowing prior knowledge about the domain and data to be included in a relatively easy and natural framework.

In addition to DM, techniques originating in AI have focused almost exclusively on dealing with data at the symbolic (categorical) level, with little attention paid to continuous variables. In machine learning and case-based reasoning, algorithms for classification and clustering have focused heavily on heuristic search and nonparametric models. Emphasis on mathematical rigor and analysis of results has not been as strong as in statistics or pattern recognition, with the exception of computational learning theory, which has focused on formal general worst-case bounds for a wide class of representations [24, 31]. Machine learning work contributes mainly to the DM step of the process, with some contributions in the area of representation and selection of variables through significant search.

DM does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions [14, 7].

Statistics plays an important role primarily in data selection and sampling, DM, and evaluation of extracted knowledge steps. Historically, most statistics work has focused on evaluation of model fit to data land on hypothesis testing. These are clearly relevant to evaluation the results of DM to filter the good from the bad, as well as within the DM step itself in searching for, parametrizing, and fitting models to data [30, 31].

On the limitations front, work in statistics has focused mostly on theoretical aspects of techniques and models. Thus, most work focuses on linear models, additive Gaussian noise models, parameter estimation, and parametric methods for a restricted class of models. Search has received little emphasis, with emphasis on closed-form analytical solutions whenever possible. While the latter is very desirable both computationally and theoretically, in many practical situations a user might not have the necessary background statistics knowledge (which can often be substantial) to appropriately use and apply the methods. Furthermore, the typical require an Apriori model and significant domain knowledge of the data as well as of the underlying mathematics for proper use and interpretation [31].

Finally, the key point is that DM is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. DM is a tool for increasing the productivity of people trying to build predictive models [31]. Therefore, the researcher applied DM in order to predict the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS area.

2.6 Data Mining Process

In the previous section the researcher had explained the difference between DM and KDD. Before presenting the state of the art of the existing processes, we mention that in the scientific literature there is a lot of confusion between the terms "process" and "methodology". A process is represented by a sequence of steps executed in order to produce a certain result. A methodology is defined as an instance of a process, by specifying the tasks that should be executed, the inputs, the outputs and the way the tasks should be executed. In brief, a process gives the user the tasks that should be executed and a methodology tells the user also "how to" perform those tasks [57].

As the researcher said before DM is one among the most important steps in the KDD process. It can be considered the heart of the KDD process. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the dataset [57].

Nowadays, with the explosion of information, DM has become one of the top ten emerging technologies that will change the world [8].There are two basic styles of DM: hypothesis testing and KD. Hypothesis testing is a top-down approach that is used when a confirmation or a rejection of an already defined hypothesis is needed. The other style is KD (relevant for this research study). It is a bottom-up approach and it is used when we want to find something that we do not know searching available data. It can be directed or undirected. There is no target field in undirected knowledge discovery. Instead, what the researcher wants from a computer is to recognize the schemes within the data that are of some importance [32].

DM is a rather complicated process that has to be planned very carefully in order to be successful. It has to be organized within one of the proposed rigorous procedures [24].

According to Sumathi and Saarenvitra [30],”once a data warehouse has been developed, the DM process falls into four basic steps: data selection, data transformation, DM, and result interpretation”.

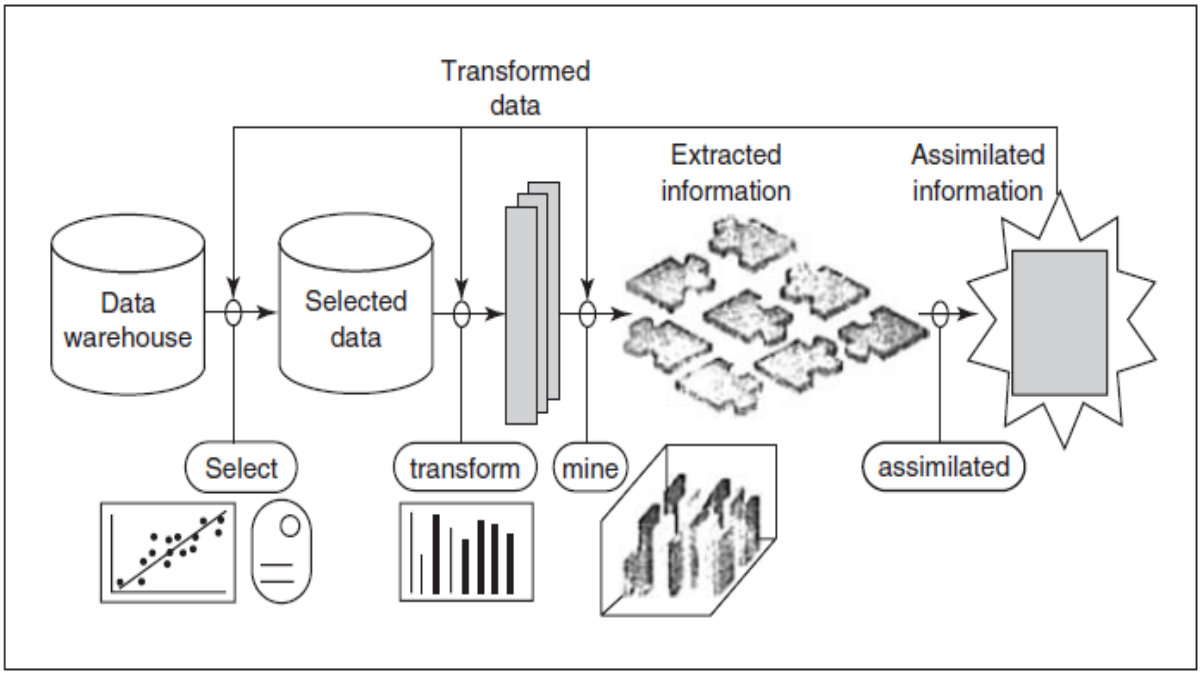


Figure 2.1: Data Mining Process

The DM process is iterative and interactive. The process is iterative, which means that sometimes it may be necessary to repeat the previous steps. The problem with this process, as with all the existing processes for DM, is the lack of user guidance.

2.6.1 Identifying the Target Dataset (Selection)

The first step in the DM process is to select the target data. The selected data types may be organized along multiple tables: during data selection, the user might need to perform table joints. Furthermore, even after selecting the desired database tables, mining the contents of the entire table is not always necessary for identifying useful information. Under certain conditions and for certain types of DM operations (such as when creating a classification or regression model), it is usually a less expensive operation to sample the appropriate table, which might have been created by joining other tables, and then mine only the sample. Therefore, in this research study the data was selected from the BRHP DSS dataset.

2.6.2 Preparing the Data for Analysis (Transformation)

After selecting the desired database tables and identifying the data to be mined, the user typically needs to perform certain transformations on the data. Three considerations dictate which

transformation to use: the task (mailing-list creation, for example), the DM operations (such as predictive modeling), and the DM technique (such as decision trees) involved. Transformation methods include organizing data in desired ways (organization of individual client data by household), and converting one type of data to another (changing numeric values into nominal ones, so, that they can be processed by a decision tree). Another transformation type, the definition of new attributes (derived attributes), involves applying mathematical or logical operators on the values of one or more data base attributes-for example, by defining the ratio of two attributes [18]. So, the researcher transformed some BRHP's attributes by deriving attributes.

2.6.3 Building and Testing the Model (Data Mining)

The user consequently mines the transformed data using one or more techniques to extract the desired type of information. For example, to develop an accurate, symbolic classification model that predicts whether magazine subscribers will renew their subscriptions, a circulations manager might need to first use clustering to segment the subscriber database, and then apply rule induction to automatically create a classification model for each desired cluster. The problem of overfitting is another issue that deserves a due consideration at this step. According to Mitchell [27], "overfitting is an attempt to create overly complex DM model that fits noise in the training data or unrepresentative features of the particular training data that decreases the generalization accuracy of the model over other unseen instances". Therefore, the researcher tried to build the predictive model by pruning the tree in order to prevent the overfitting situation and also used several experiments to get the best fitted.

2.6.4 Evaluating the Model (Result Interpretation)

The user must finally analyze the mined information according to his decision-support task and goals. Such analysis identifies the best of the information. For example, if a classification model has been developed, during result interpretation, the DM application will test the model's robustness, using established error-estimation methods such as cross validation. During this step, the user must also determine how best to present the selected mining-operation results to the decision maker, who will apply them in taking specific actions. For example, the user might decide that the best way to present the classification model is logically in the form of if-then

rules. So, the investigator tried to evaluate the model by using model testing with confusion matrix performance method.

Three observations emerge from this four-step process:

- Mining is only one step in the overall process.
- The process is not linear but involves a variety of feedback loops.
- Visualization plays an important role in the various steps.

2.7 Data Mining Techniques

DM techniques provide people with new power to research and to manipulate the existing large volume of data. The goal of DM can be varied based on the intended use of the system. Different methods and techniques are needed to find different kinds of patterns. Health care now collects data in gigabytes per hour volume. DM can help with data reduction, exploration, and hypothesis formulation to find new patterns and information in data that surpass human information processing limitations [18, 59].

As mentioned earlier, there is a proliferation of reports and articles that apply DM and KDD to a wide variety of health care problems and clinical domains and includes diverse projects related to cardiology, cancer, diabetes, finding medication errors, and many others. Within DM methodologies, one may select from an extensive array of techniques that include, among many others, classification, clustering, and association rules [59].

There are two high-level primary goals of DM; they are predictive and descriptive models. Predictive mining tasks perform inference on the current data for seek of predicting the target that the user interested on it. Descriptive mining involves on presenting patterns that describe the data in the form that could be understood by the user. In the context of KDD, description tends to be more important than prediction [22]. In this study, a DM model and a proper DM implementation was achieved in a BRHP DSS database system.

2.7.1 Predictive Modeling

At the heart of predictive model, DM is the process of building a model to represent the data set and to carry out the DM operation. Predictive modeling has the specific objective of allowing us

to predict the value of some target characteristic of an object based on observed values of other characteristics of the object [12].

Moreover, predictive modeling is a technique that involves using some variables or fields in the dataset to predict unknown or future values of other variable of interest. It used to develop a model to relate a dependent variable with a set of independent variables. Historical data is employed in order to build and train a model that describes the current observed behaviour [28].

In this paper, the researcher tried to show implementation on the predictive model, since this part of the project is one which is more portable across BRHP DSS dataset to predict the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS area.

2.7.1.1 Classification Techniques

The rule inductions from J48 decision tree and Naïve Bayes classifier DM techniques used in the research project fall under the category of machine learning that uses high end modeling techniques for uncovering hidden patterns and/or predicting outcomes of the child survival in order to predict the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS sites.

To be useful for DM purposes, it creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service of epidemiological surveys. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent to an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database to predict patterns of under-fives mortality in Ethiopia [7, 31].

In a follow-up study, classification problems aim to identify the determinants that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave [31].

In addition to the above listed research efforts, there are other studies related to using DM for prediction in medical domains. Supervised induction process involves automatically creating a classification model from a set of records (examples) called the training set. The training set might either be a sample of the database or a warehouse being mined, the entire database, or a data warehouse. The records in the training set must belong to a small set of classes that the analyst has predefined. The induced model consists of patterns-essentially generalizations over the records that are useful for distinguishing the classes. Once induced, a model can help automatically predict the class of other unclassified records. Supervised induction methods can be either neural or symbolic. Neural methods such as back propagation represent the model as architecture of nodes and weighted links [18]. Symbolic methods create models that are represented either as decision trees or as if-then rules. A supervised induction technique is particularly suitable for DM if it has three characteristics [30].

A. It can produce high-quality models even when the data in the training set is noisy and incomplete.
B. The resulting models are comprehensible and explainable, so that, the user can understand how the system makes the decision.
C. It can accept domain knowledge, which can expedite the induction task while simultaneously improving the quality of the induced model.

2.7.1.1.1 Classification by Decision Tree Induction

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [7].

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore, is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast [7, 30].

Decision trees are especially attractive in DM environments since human analysts readily comprehend the resulting models. A record can be associated with a unique leaf node by starting at the root and repeatedly choosing a child node based on the splitting criterion, which evaluates a condition on the input records at the current node [31].

Moreover, during tree construction and attribute selection measures are used to select the attribute that best partitions the tuples into distinct classes. When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. Decision tree construction algorithms consist of two stages: tree building and pruning. In the former, most decision tree construction algorithms grow the tree top down in the following greedy way. Starting with the root node, the database is examined by "split selection method" for selecting the split condition at each node. The database is then partitioned and the procedure applied recursively. In the pruning stage, the tree constructed in the tree building phase is pruned to control its size, and sophisticated pruning methods select the tree in a way that minimizes prediction errors [30].

In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems [7].

As Rea [11] puts it "once classes are defined the system should enter rules that govern the classification, therefore, the system should be able to find the description of each class". The writer further explains that the descriptions should only refer to the predicting attributes of the training set.

Decision Tree is known to have an advantage over numerous techniques due to the output it produces. The output of a decision tree is transparent, which makes it easy for users or non technical persons to understand. However, decision tree techniques are known to have scalability and efficiency problems, such as substantial decrease in performance and poor use of available system resources [46]. For this research project, a classification task is to be carried out since a

model is to be built by using the pre-classified data of past records of children that are included in the study base of the BRHP DSS dataset.

2.7.2 Descriptive Model

One can easily note that a descriptive model presents the main features of the data. It is essentially a summary of the data, permitting us to study the most important aspects of the data without their being obscured by the sheer size of the data set [12]. It is finding human-interpretable patterns, associations or correlations describing the data.

Sometimes, the DM problem is simply to describe what is happening in a complicated database. This includes knowing the people the products or process that are applicable and which constitute the database [24]. The distinction between description and prediction is not very sharp. Predictive models can also be descriptive (to the degree that they are understandable), and descriptive models can be used for prediction [29].

2.7.2.1 Clustering Techniques

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so, may be considered as a form of data compression [11]. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Clustering has its roots in many areas, including DM, statistics, biology, and machine learning [7].

Cluster analysis is an important human activity. Early in childhood, as the investigators inscribe that learn how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. By automated clustering, As the writer can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing [30, 31]. Cluster analysis takes ungrouped data and uses automatic techniques

to put this data into groups. Clustering is unsupervised, and does not require a learning set. It shares a common methodological ground with classification. In other words, most of the mathematical models mentioned earlier in regards to classification can be applied to cluster analysis as well.

There are several clustering techniques, organized into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data (such as frequent pattern based methods), and constraint-based clustering. Clustering can also be used for outlier detection [7]. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS [30].

In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. Data clustering is under vigorous development. Contributing areas of research include DM, statistics, machine learning, spatial database technology, biology, and marketing. Owing to the huge amounts of data collected in databases, cluster analysis has recently become a highly active topic in DM research [7, 36].

Two Crows Corporation [31] mentioned that the goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. The categories (clusters) can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories. According to Han and Kamber [7], each cluster that is formed can be viewed as a class of objects from which rules can be derived.

Other research on scalable clustering focuses on training databases with large attribute sets. The search methods involve discovering the appropriate subspace of attributes in which the clusters are most likely to exist. These methods help analysis trying to understand the results, as they focus only on the attributes associated with a given cluster [30].

2.7.2.2 Link Analysis

Link analysis is a descriptive approach to exploring data that can help identify relationships among values in a database. The two most common approaches to link analysis are association

discovery and sequence discovery. Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery. Sequence discovery is very similar, in that a sequence is an association related over time [30, 36].

The problem of detecting patterns was first introduced in the application domain of market basket analysis to find association between two sets of bought products. Thus, initial research was concentrated on the discovery of Boolean association rules. However, more recent work is focusing on quantitative association rules [7].

As mentioned earlier, link analysis is concerned with finding rules between data elements. The two most common rules are association and sequencing rules. According to Shegaw [11], association methods discover rules of the form: "if item A is part of an event, then X% of the time item B is also part of the event".

Associations are written as A & B, where A is called the antecedent or left-hand side (LHS), and B is called the consequent or right-hand side (RHS). For example, in the association rule "If people buy a hammer then they buy nails," the antecedent is "buy a hammer" and the consequent is "buy nails" [31].

It's easy to determine the proportion of transactions that contain a particular item or item set: simply count them. The frequency with which a particular association (e.g., the item set "hammers and nails") appears in the database is called its support or prevalence. If, say, 15 transactions out of 1,000 consist of "hammer and nails," the support for this association would be 1.5%. A low level of support (say, one transaction out of a million) may indicate that the particular association isn't very important -or it may indicate the presence of bad data (e.g., "male and pregnant") [31].

To discover meaningful rules, however, as the investigator must also look at the relative frequency of occurrence of the items and their combinations. Given the occurrence of item A (the antecedent), how often does item B (the consequent) occur? That is, what is the conditional predictability of B, given A? Using the above example, this would mean asking "When people buy a hammer, how often do they also buy nails?" Another term for this conditional

predictability is confidence. Confidence is calculated as a ratio: (frequency of A and B)/ (frequency of A) [31].

Moreover, association rules capture the set of significant correlation's present in a given data set. Given a set of transactions, where each transaction is a set of items, an association rule is an implication of the form X => Y, where X & Y are sets of items. This rule has support *s* if *s%* of transactions includes all the items in both X and Y, and confidence *c* if *c%* of transactions containing X also contain Y. For example, the rule "[carbonated beverages] and [crackers]=>[milk]" might hold in a supermarket database with 5% support and 70% confidence [30].

Finally, the goal is to discover all association rules with support and confidence greater than the user-specified minimum support and minimum confidence, respectively. This formulation has been extended in many directions, including the incorporation of taxonomies, quantitative associations, and sequential patterns [30].

2.8 Data Mining Methodologies

The ultimate goal of the KD Process (henceforth KDP) model is to achieve overall integration of the entire process with industrial standards. Another important objective is to provide interoperability and compatibility between the different software systems and platforms used throughout the process. Integrated and interoperable models would serve the end user in automating, or more realistically semi-automating, work with KD systems. The efforts to establish a KDP model were initiated in academia. [63].

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

2.8.1 Knowledge Discovery in Database (KDD)

As mentioned earlier, KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular DM method. While the DM step is characterized by the extraction of patterns hidden in the data, the whole KDD process is broader and includes all the processing (data selection, preprocessing and transformation) that is needed for this to occur, making it possible to evaluate and interpret the results that were obtained after the DM techniques were used. The KDD process is a set of continuous activities that include five steps: Data Selection, Preprocessing, Formatting, Data Mining and Interpretation [14, 35 and19].

Moreover, the process starts by understanding the application's domain and the targets that must be reached. Then, a selection can be drawn from these data so that, one may work with the data that are of interest. The pre-processing step is the one in which missing or inconsistent data are analyzed and treated. During the formatting step data are prepared. So, DM can be used as, for instance, to map categorical data among numerical data or using methods to reduce dimensions in the data [22, 19]. According to Silver [60], "pre-processing and formatting may take up to 80% of the time needed for the whole process".

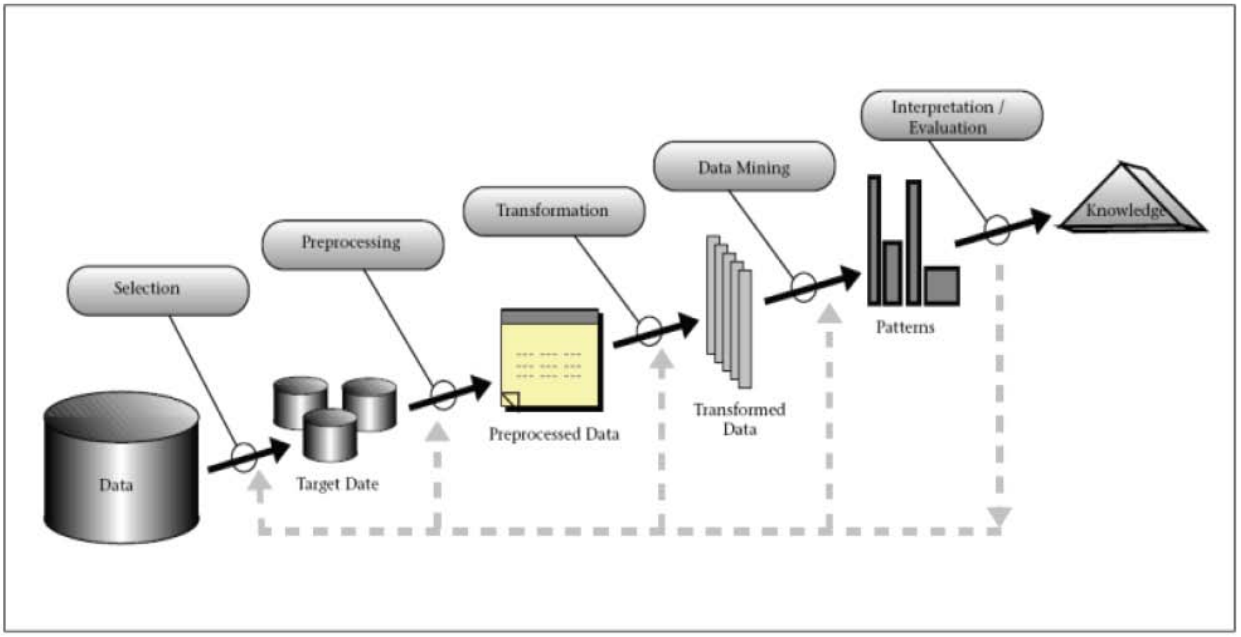


Figure 2.2: KDD Process model

The main advantage of the knowledge discovery process is that no hypotheses are needed and knowledge is extracted from the data without previous knowledge. KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular DM method. The KDD process is a set of continuous activities that include five steps: Data Selection, Pre-processing, Formatting, DM, and Interpretation. The main purpose with the KDD process is to obtain knowledge hidden in data that may be useful for decision-making, by using methods, algorithms, and techniques from different scientific areas. According to Tan et. al. [60], these include "Statistics, Artificial Intelligence, Machine Learning, and Pattern Recognition" [35, 19].

The KDD process consists of the following five steps:

1) **Select a target data set:** The initial step is based on data needed for the DM process may be obtained from many different and heterogeneous data sources.
2) **Data preprocessing:** In this step the data to be used by the process may have incorrect or missing data. There may be abnormal data from multiple sources involving different data types and metrics.
3) **Data transformation:** Attributes and instances are added and/or eliminated from the target data. Data from different sources must be converted into a common format for processing.
4) **Data mining:** A best model for representing the data is created by applying one or more DM algorithms.
5) **Interpretation/evaluation:** The final step the researcher examines the output from step 4 to determine if what has been discovered is both useful and interesting.

Another important step not contained in the KDD process is goal identification. The focus of this step is on understanding the domain being considered for KD.

2.8.2 CRoss Industry Standard Process for Data Mining (CRISP-DM)

The goal of the project was to define and validate an industry- and tool-neutral DM process model that which would make the development of large as well as small DM projects faster, cheaper, more reliable and more manageable.

CRISP-DM is a process model because it is the “de facto standard” for developing DM and KD projects. In addition, CRISP-DM is the most used methodology for developing DM projects. Analyzing the problems of DM and KD projects, a group of prominent enterprises developing DM projects, proposed a reference guide to develop DM & KD projects. This guide is called CRISP-DM or independent so it can be used with any DM tool and it can be applied to solve any DM problem. CRISP-DM defines the phases to be carried out in a DM project as well as defines for each phase the tasks and the deliverables for each task [19,35].

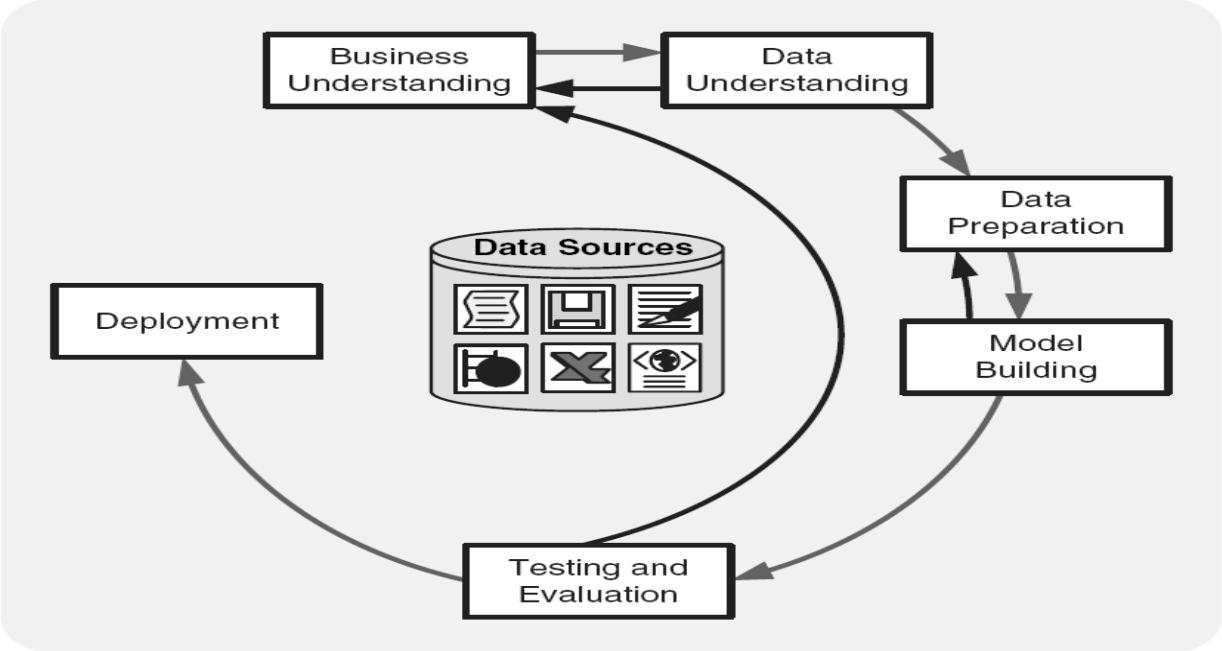


Figure 2.3: CRISP-DM Process model

A. **Business understanding**: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

B. **Data understanding**: This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

C. **Data preparation**: The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.

D. **Modeling**: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type.

E. **Evaluation**: Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives.

F. **Deployment**: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.8.3 SEMMA

The methodology and approach that SAS Institute proposes is referred to as SEMMA, for Sample, Explore, Modify, Model, and Assess. Beginning with a statistically representative sample of data, users can apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and affirm the model's accuracy [22, 19].

A. **Sample:** The first step is to extract a portion of a large data set big enough to contain the significant information yet small enough to manipulate quickly.

B. **Explore:** This phase involves searching speculatively for unanticipated trends and anomalies so as to gain understanding and ideas.

C. **Modify:** The insights that are gained from the exploration phase enable knowledge workers to group the most productive subsets and clusters of data together for further analysis and exploration.

D. **Model:** This process involves searching automatically for a variable combination that reliably predicts a desired outcome.
E. **Assess:** During this evaluation process, assessment of the results gained from modeling provides indications as to which results should be conveyed to senior management, how to model new questions that have been raised by the previous results and thus, proceed back to the exploration phase.

2.8.4 Hybrid Model

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model developed by Cios et. al. It was developed based on the CRISP-DM model by adopting it to academic research [14, 19 and 35].

A hybrid of the above-mentioned approaches may be considered in determining a suitable goal for DM. All KDD model process models emphasise the iterative nature of the process that a DM application is conducted. Typically, goals are selected, an experiment is conducted, based on results at each stage, a step is revisited or moves to the next step [19].

The iterative nature of KDD model process models allows retracting and considering different approaches/paths (goals, techniques and methods) in conducting a DM experiment as a way to address this uncertainty. This approach however is not optimal and results in a trial-and-error process, which is resource-intensive and risky with no guarantee of favourable results. Approaches to minimise unsuccessful attempts and provide certain guarantees would be highly beneficial [35, 19].

A description of the six steps follows:

1. **Understanding of the problem domain:** This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem.

2. **Understanding of the data:** This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts.

3. **Preparation of the data:** This step concerns deciding which data will be used as input for DM methods in the subsequent step.

4. **Data mining:** Here the data miner uses various DM methods to derive knowledge from preprocessed data.

5. **Evaluation of the discovered knowledge:** Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

6. **Use of the discovered knowledge:** This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains.

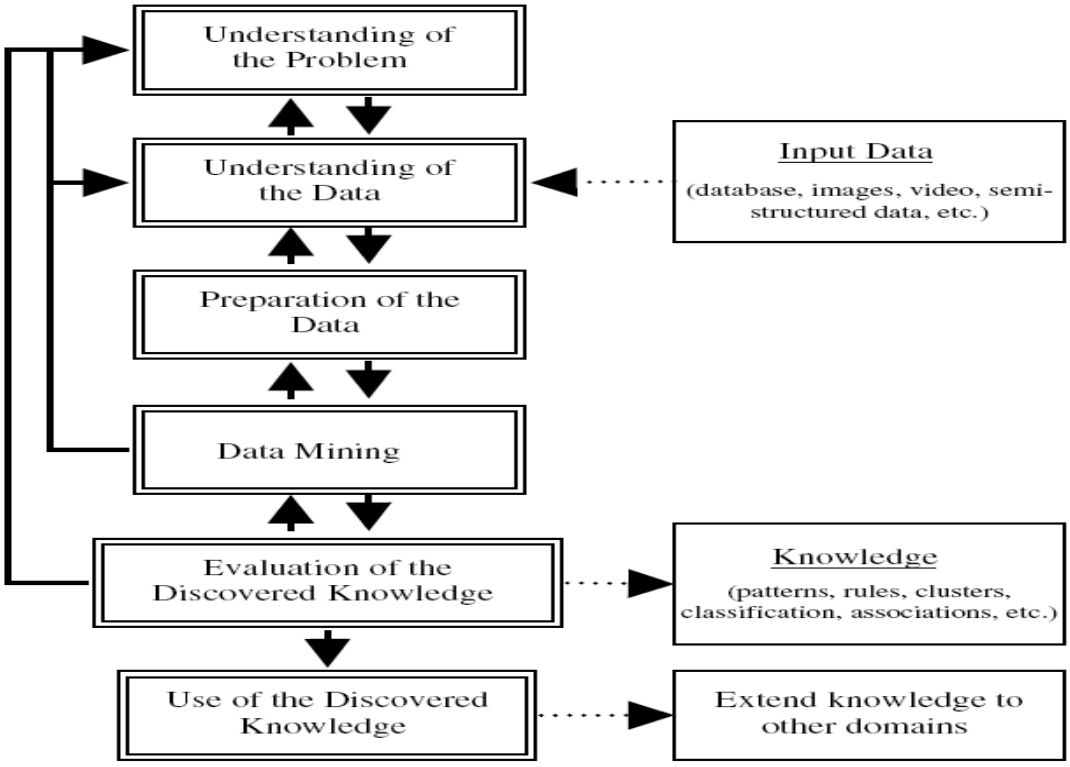


Figure 2.4: Hybrid Process model

In this study the investigator selected the hybrid process model to predict the pattern of under-five mortality in Ethiopia, particularly in BRHP DSS area, since the model emphasizes the

iterative aspects of the process, drawing from the experience of users of previous models. It identifies and describes several explicit feedback loops.

2.9 Application of Data Mining in Child Healthcare

DM applications can greatly benefit all parties involved in the healthcare industry. For example, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The huge amounts of data generated by healthcare transactions are too complex and huge to be processed and analyzed by traditional methods. Predictive analytics include a variety of techniques from DM that analyze the statistical data to make predictions about future events [25].

In addition, every medical treatment has risk that must be compared to the risk of the disease. It is important to calculate the actual risk rather than to make treatment recommendations based upon perceived risk [25].

A lot of valuable knowledge hidden in the database can be found using DM approach, which is worth exploring. For example, using the large item sets and decision tree classification method in DM technique to infer the relationships between characteristics of patient symptoms and the illnesses ,so that, patients can utilize the results of this research to assist in guiding patients to connect their own symptoms to the type of illness accurately[25]. Using the cluster technique in DM to discuss the clustering model of individual physician service time for patient's treatment can be efficient. This model classifies different patient groups, their corresponding population ratios, and physician service time on each different group according to similarity of attributes [25].

DM technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Following are some of the important areas of interests where DM techniques can be of tremendous use in health care management [58].

- Data modeling for health care applications
- Executive Information System for health care
- Forecasting treatment costs and demand of resources
- Anticipating patient's future behavior given their history
- Public Health Informatics
- E-governance structures in health care
- Health Insurance

Traditionally, decision making in health care is based on the ground information, lessons learnt in the past resources and funds constraints. However, DM techniques and knowledge management technology can be applied to create knowledge rich health care environment.

Finally, healthcare data is massive. It includes patient centric data, resource management data and transformed data. Healthcare organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and DM techniques may help in answering several important and critical questions related to health care [58].

2.10 Related Works

Shegaw [11] to predict child mortality patterns on the application of DM technology: the case of BRHP. A data set totaling 1,100 records of children was used to build and test both neural network and decision tree models. In order to build models that can predict the risk of child mortality, several models were built by employing both neural network and decision tree approaches. The best performing neural network model and decision tree classifier were then chosen and evaluated using ten previously unseen records of children.

The methodology employed consisted of three basic steps; data collection, data preparation, and model building and testing. However, since a DM task is an iterative process, these steps were not followed strictly in linear order. There were instances where there was a need to go back and forth between the different steps.

Using the neural network approach, the best model was identified for the training made by using the default parameters (i.e. training tolerance of 0.1, learning rate of 1.0, and smoothing factor of 0.9) and the following 9 input variables: "ENVIRN", "AGE", "OUTMIG", "HHRELIG",

“HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, AND “WINDOWS”. This model had an accuracy rate of 93% (classified 102 of the 110 test cases correct) at a testing tolerance of 0.4 and was tested with accuracy of 88 % (classified 97 of the 110 test cases correct) at testing tolerances of 0.2 and 0.1.

The decision tree classifier was built by using the following attributes:“ENVIRN”,“AGE”, “SEX”,“OUTMIG”,“HHRELIG”,“HHETHNIC”,“HHLITERAC”,“HHHEALHTH”,“HHWATE R”,”HHMEMBAVE”, “HHLIVESTOK”, AND “WINDOWS”. This classifier resulted with an accuracy of 95% (i.e. it classified 942 of the 995 training cases correct) on training cases and it achieved 95% accuracy (classified 105 of the 111 test cases correct) on test cases.

The results obtained in this research work have proved the potential applicability of DM technology to predict child mortality patterns based solely on demographic, parental, environmental, and epidemiological factors. The encouraging results obtained from both neural networks and decision trees indicate that DM is really a technology that should be considered to support child health care prevention and control activities at the district of BRHP DSS area and at a national level in general.

In general, encouraging results were obtained by employing both neural networks and decision tree approaches. Although both neural network and decision trees showed comparable accuracy and performance in predicting the risk of child mortality, the decision tree approach seems more applicable and appropriate to the problem domain.

Abera [13] to determine the child mortality in BRHP DSS by using retrospective cohort study methods, this study was a retrospective cohort study that took secondary data of BRHP and qualitative study design to supplement on the quality of data collection. All birth cohorts born between Jan 1st to Dec 31st, 2000 were considered as the study population. Data was analyzed using the Cox proportional Hazard model to track survival pattern of children and factors associated with child death. Results: Infant and under five mortality rates were 83.9 and 118 deaths per 1000 live births. Excess mortality was observed in female children than in males; moreover, multiple births were at increased risk of dying than singleton. Urban children had more (50%) chances of survival compared to rural ones. upon stepwise multivariate Cox regression source of water esp. pipe water, sex of child, multiple births, urban places of residence

and availability of radio in the household were found to be independent predictors of child survival. Finally, the researcher concluded that mortality is relatively high and the provision of safe and adequate water supply and promotion of child health should be considered in the area.

Taddesse [17] to mine vital statistics data by using the application of DM technology: the case of BRHP. In his research, he used the BRHP database as the experimental study that consists of 25 attributes and 66,123 cases sampled (95,220 cases after SMOTE) from 236,549 cases.

He used J48 decision tree algorithm .The main aim of the research was to identify the best performing scenario of DM, the technique with knowing the most determining factor/attribute for the given dataset of the research.

Several models were developed as experimental analysis to outperform some of the J48 parameters. The models built allow as more flexible with our output and can be more powerful weapons in our DM arsenal. As the investigator can see from experiment 90.3% predictive accuracy was obtained for the selected best model .That means 90.3% of the test data represents the majority class of the training set. The time required for computation and classification in this method is minimal.

The prediction rate of the J48 decision tree algorithm had revealed that mining the vital statistics data in BRHP is possible or applicable with 90% accuracy. The result shows that using the SMOTE approach can improve the accuracy of classifiers for a dataset.

Hence, it was possible to conclude that the vital statistics data (death or mortality dataset) can be predicted by the application of classification technique (J48 decision tree algorithm) given the limitation of this study.

CHAPTER THREE

METHODS FOR MINING UNDER-FIVES MORTALITY DSS DATA

As previously expressed in the methodology section in chapter one, the focus of this thesis is on realizing predictive models from knowledge of classification. The aim is to model connections between input, or predictor variables, and the outcome or prediction using observed classification. Hence, it is important to explain the classification implementation for model building and experiments were carried out in the DM process, which also involve DM tool selection and algorithms were used for modeling. The DM algorithms used in this research was to predict the pattern of under-five mortality in Ethiopia based on BRHP DSS datasets were J48 decision tree algorithm and Naïve Bayes classifier.

There are a number of machine intelligent tools, techniques and algorithms that are available in the market but at the same time not all tools, techniques and algorithms are the best for all problems in the dataset. Different data sets will produce different results based on the algorithms used. In this thesis the researcher tested some algorithms based on decision trees rule based classification and Naïve Bayes classifier probability. The researcher's aim was to find the best tool, techniques, and algorithms to predict the pattern of under-five mortality in Ethiopia that based on the available data in the BRHP DSS dataset.

3.1 The WEKA Tool

WEKA was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems-and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset. WEKA is a collection of machine learning algorithms for solving real-world DM problems. It contains 41 different algorithms for classification and numeric prediction [52].

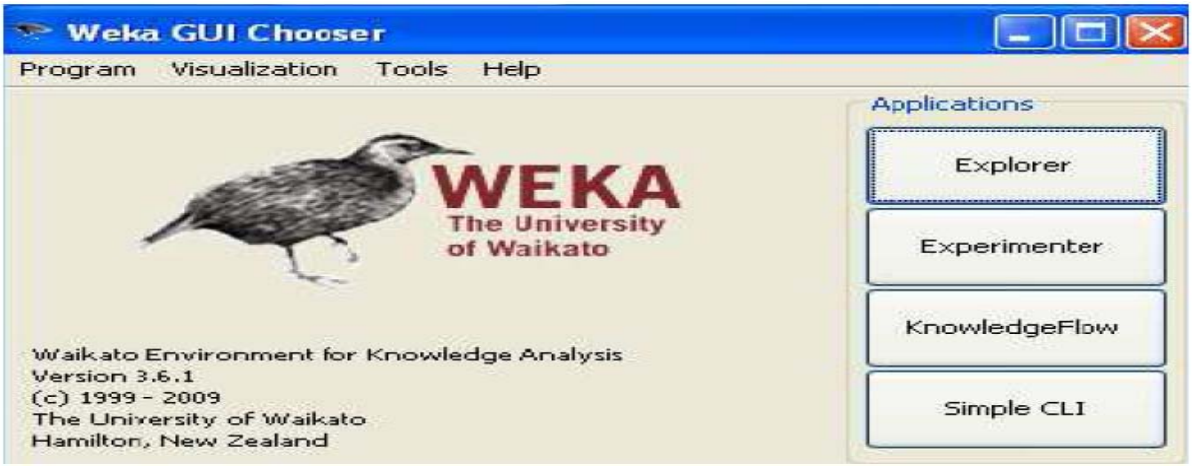


Figure 3.1: WEKA GUI application main window

A number of DM methods were implemented and experimented in the WEKA software. Some of them were based on decision trees like the J48 decision tree, some are rule-based like PART and decision tables, and some of them are based on probability and regression, like the Naïve Bayes algorithms were implemented.

One way of using WEKA is to apply a learning method to a dataset and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances. A third is to apply several different learners and compare their performance in order to choose one for prediction. The learning methods are called classifiers, and in the interactive WEKA interface you select the one you want from a menu lists. Many classifiers have tunable parameters, which you access through a property sheet or object editor. A common evaluation module is used to measure the performance of all classifiers [52].

Implementations of actual learning schemes are the most valuable resource that WEKA provides. But tools for preprocessing the data, called filters, come a close second. Like classifiers, you select filters from a menu and tailor them to your requirements [15]. The investigator showed how different filters can be used, list the filtering algorithms, and describe their parameters in the prediction of the pattern of under-five mortality in Ethiopia, particularly in BRHP DSS area.

The data is often presented in a SPSS, Epi-info, spreadsheet or database. However, WEKA's native data storage method is attribute relation file format (henceforth ARFF) format. You can easily convert from a spreadsheet to ARFF. The bulk of an ARFF file consists of a list of the

instances, and the attribute values for each instance are separated by commas. Most spreadsheet and database programs allow you to export data into a file in comma-separated value (henceforth CSV) format as a list of records with commas between items. Having done this, you need only load the file into a text editor or word processor; add the dataset's name using the @relation tag, the attribute information using @attribute, and a @data line; and save the file as raw text. However, you don't actually have to go through these steps to create the ARFF file yourself, because the explorer can read CSV spreadsheet files directly [52].

3.2 Decision Tree Classifiers

Decision tree is powerful in its functions and a very popular tool for classification and making prediction. It is constructed top-down: specific instances are put in sets, and as the tree grows, smaller subsets which are mainly leaf and branches are gradually divided. Each leaf represents the classifying outcomes of the decision tree, and each branch from a leaf node corresponds to the possible values/criterions for this attribute. Decision trees can classify variables according to a certain rule or classify data based on some data characteristics [33].

Decision tree is one of the easier data structure to understand DM. Rules from the training dataset are first extracted to form the decision tree which is then used for classification of the testing dataset. A decision tree is necessarily a tree with an arbitrary degree that classifies instances. They are a powerful tool for classification and predication but require extensive computation. Creating the tree based on the training set takes time although making decisions once the tree is made is not time consuming. Classification tree algorithms may be divided into two groups: one whose result is a binary tree and other that yields non-binary trees (also called multiway) splits [56].

In decision trees, the leaf node represents the complete classification of a given instance of the attribute and the decision node specifies the test that is conducted to produce the leaf node. Thus, with a decision tree, the sub tree that is created after any node is necessarily the outcome of the test that was conducted.

In addition, a decision tree is used to classify a certain instance from the root of the tree till the leaf node which provides the outcome of that instance. A major issue in using decision tree is to

find out how deep the tree should grow and when it should stop. Usually, if all the attributes are different and lead to the same outcome, the decision tree might not be the most effective in making decision and, at the same time, the size of the tree will be large [50].

There are a number of algorithms that are based on decision trees. The investigator compared results of different decision tree based tools, techniques and algorithms to evaluate each for a given dataset. The researcher hopes to determine the decision tree or algorithm that provides better accuracy for the particular dataset. Some of the most common and effective types of algorithms based on decision trees are C4.5, PART and CART [56].

The C4.5 algorithm is a part of the multiway split decision tree. C4.5 yields a binary split if the selected attribute/variable is numerical, but if there are other variables representing the attributes it will result in a categorical split. That is, the node will be split into C nodes where C is the number of categories for that attribute [56]. The J48 decision tree in WEKA is based on the C4.5 decision tree algorithm.

The presence of nodes containing insignificant attributes results in increased depth, which detracts from the effectiveness of decision trees. Moreover, decision tree algorithms not only choose the best splitting attribute for a node, but also decide what values or how many branches to assign to that node. Poorly designed decision tree algorithms that assign random values often cause an ineffective rendering of the decision tree technique [56].

The C4.5 algorithm: utilizes the same basic inductive tree creation approach as ID3, but extends its capabilities to classification of continuous data by grouping together discrete values of an attribute into subsets or ranges. Another advantage of C4.5 is that it can predict values for data with missing attributes based on knowledge of the relevant domains [50].

As mentioned above, the researcher implemented C4.5 learners, each based on J48, but with the possibility of using one of three alternative models in the leaves to construct the best fitted model:

- **J48-Linear:** each leaf may contain a classifier that uses linear regression functions to approximate class membership (the so-called Classification via Regression classifier in WEKA [44]).

- **J48-IB1:** a leaf may contain a simple nearest neighbor classifier [39] using one neighbor (i.e., IB 1, in the terminology of [38]).
- **J48-Bayes:** a leaf may contain a Naïve Bayes Classifier [41] that uses a normal distribution assumption for the continuous attributes [40].

The learning algorithm was presented with a set of examples relevant to the problem domain classification task done at the BRHP DSS dataset in order to predict the pattern of under-five mortality. The aim of the learning method is to produce a tree that correctly classifies all examples in a subset of the training set. All other examples in the training set are then classified using the tree. If the tree gives the correct answer for all of these examples then it is correct for the entire training set, and the iterative process terminates. If not, a selection of the incorrectly classified examples is added to the initial subset and the process starts again [43].

Moreover, the researcher implemented a divide-and-conquer strategy was used to construct the proposed decision tree [42]. The choice of the test option mode to partition the training set was crucial for the complexity of the inducted tree. The test is to select an attribute for the root tree and subsequent sub-trees with the parameter of test option of pruned or unpruned situation.

The information gain measure was used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that simple (but not necessarily the simplest) tree is found. So, after the above proposed procedures the following explanations were used for the BRHP DSS database to formulate the best attributes selection by using entropy and information gain techniques.

In addition to select the test attribute (i.e., the best attribute for splitting), the entropy and information gain need to be calculated for each attribute. The C4.5 algorithm adopts an information-based method that relies on two assumptions [29]. If set S represent the training set

and x, y and z are number of examples of classes X, Y and Z respectively, then the assumptions are used to formulate the best attributes selection that can be applied on the WEKA tool:

- Any correct decision tree for S will classify examples in the same proportion as their representation in S. Thus, an arbitrary example belongs to class X, Y and Z with probability:

$$\frac{x}{(x+y+z)}, \frac{y}{(x+y+z)} \text{ or } \frac{z}{(x+y+z)} \text{ respectively and}$$

- When a decision tree is used to classify an example (training set), it returns a class. A decision tree can thus be regarded as a source of a message X, Y, or Z, with the expected information needed to generate this message given by:

$$I(X,Y,Z) = -\frac{x}{(x+y+z)}\log_2\left(\frac{x}{(x+y+z)}\right) - \frac{y}{(x+y+z)}\log_2\left(\frac{x}{(x+y+z)}\right) - \frac{z}{(x+y+z)}\log_2\left(\frac{z}{(x+y+z)}\right)$$

From these assumptions, the expected information required for the decision tree with attribute A as its root is given by:

$$E(A) = \sum \frac{x_i + y_i + z_i}{x+y+z} \cdot I(x_i + y_i + z_i)$$

Where x_i, y_i, and z_i are number of examples for the classes of X, Y and Z respectively with value A_i of the attribute A. The summation gives the total expected information for attribute A. The information gained by branch the tree on A is:

$$\text{GAIN (A)} = I(X, Y, Z) - E(A)$$

At each non-leaf node of the decision tree, the gain of each untested attribute is determined. This gain in turn depends on the value of x_i, y_i, and z_i for each value A_i of the attribute A. Every example is examined to determine its class and its value of A. Thus, the total computational requirement per iteration is proportional to the product of size of the training set, the number of attributes, and the number of non-leaf nodes in the decision tree. The training stage of the algorithm results is a classifier in a form of decision tree, which can be used to classify an unseen set of testing samples.

The attribute with the highest information gain is considered as the most discriminating attribute of the set under consideration. So, an attribute that yields maximum information gain will be chosen for data set partitioning. Then, a node is created and labeled with the chosen attribute, branches are formed for each value of the attribute, and the samples are partitioned accordingly. The same criteria were applied to each split sample on this research project. The iterative divide and conquer process executes until no further split was required [20].

Furthermore, a set of classification rules can be extracted from the decision tree by tracing the path from the root to each leaf (corresponding class). This set of rules can be consequently plugged into propitiate knowledge-based system [29]. So, the researcher computed the C4.5 algorithm using J48 method in order to get the best fitted model that can appropriate to predict the pattern of under-five mortality in Ethiopia, specifically for the BRHP DSS and also the investigator tried to generate rules from the J48 decision tree with comparing to the PART rules by the parameter of accuracy measures.

3.3 Naïve Bayes Classifiers

Naïve Bayes classifiers method is based on probabilistic knowledge. This method goes by the name Naïve Bayes, because it's based on Bayes's rule and "naively" assumes independence- it is only valid to multiply probabilities when the events are independent [56]. Thus, the Naïve Bayes rule outputs probabilities for the predicted class of each member of the set of test instance. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data. In Naïve Bayes classifier, the probability of the attributes are calculated based on normal distribution's mean, standard deviation, weighted sum, and precision. So, the investigator tried to show the experiments on Naïve Bayes algorithms in order to get the best fitted model for the classification as well as prediction of the BRHP DSS dataset.

Naïve Bayes gives a simple approach, with clear semantics, to representing, using, and learning probabilistic knowledge. Impressive results can be achieved using it. It has often been shown that Naïve Bayes rivals, and indeed outperforms, more sophisticated classifiers on many datasets. The moral is, always try the simple things first. Repeatedly in machine learning people have

eventually, after an extended struggle, obtained good results using sophisticated learning methods only to discover years later that simple methods such as 1R and Naïve Bayes do just as well-or even better. The Naïve Bayes classification reads a set of examples from the training set and uses the Bayes theorem to estimate the probabilities of all classifications. For each instance, the classification with the highest probability is chosen as the prediction class [52]. Most of the Bayesian classifiers utilize model that gives the probability of the data conditioned on the hypothesized model: P (XH, p), known as likelihood function [29].

Moreover, Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Studies comparing classification algorithms have found a simple Bayesian classifier known as the Naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases [56]. In this case the researcher tried to show in comparable performance by implementing several experiments with J48 decision tree algorithm and Naïve Bayes classifier on BRHP DSS database.

As mentioned before, Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve” [22, 7].

In this research study, the researcher made experiments based upon the Bayes approach defines the classification problem in terms of probabilities that formulated by the underneath proof. More specifically, the three main concepts required are conditional probability, Bayes Theorem, and the Bayes decision rule. The conditional probability P (A\B), which is used to define independent events [51, 52], is defined by

$$P(A|B) = \frac{P(A \cup B)}{P(B)},$$

Where P (A\B) is the probability that event A happens, given that B is observed. Similarly,

$$P(B|A) = \frac{P(A \cup B)}{P(A)},$$

Where P (B\A) is the probability that event B happens, given that A is observed. It then follows (by substitution) that

$$P(A \cap B) = P(A)P(B|A).$$

Although, the premise of Bayes Theorem starts with an initial degree of belief that an event was occur, and then with new information this degree of belief can be "updated" [51]. These two degrees are represented, respectively, by the prior probability P (A\B) and the posterior probability P (B\A), which are related by

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

Finally, the Bayes decision rule states that based on the posterior probabilities, it is possible to assign an element x to a class with the largest probability. In particular, for the classifying problem, the conditional probabilities described above can be defined as follows: let x be a data sample (vector of features) and w_i one of the possible classes [51]. Then, P ($x\backslash w_i$) is the prior probability, because it can be obtained based on prior knowledge (i.e., the distribution constructed from training data). Given class i, it specifies the probability of finding x within this class. Similarly, P ($w_i\backslash x$) is the posterior probability, because it is computed based on posterior knowledge. Given sample x, it specifies the probability that x belongs to class j. For a given x, if then x belongs to class 1, otherwise to class 2.

$$P(\omega_1|x) > P(\omega_2|x),$$

Where

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)},$$

The denominator term of equation is the overall probability of x in all the classes. For a given x, one must compute the posterior probabilities for all w_i classes, then assign x to the class that yields the maximum posterior probability [38].

Generally speaking, the Bayesian methodology for classification as well as prediction of the pattern of under-five mortality in Ethiopia, particularly for the BRHP DSS database follows these five steps [40]:

1. Collect data, and estimate parameters such as mean and covariance for each class (for the parametric approach the researcher assumed that all the probability density functions have a Gaussian behavior).

2. Choose a set of features.

3. Choose a model and derive a decision rule with these parameters.

4. Train the classifier and apply the decision rule by using a discriminant function (a way to represent a pattern classifier), and apply it to a test data set to classify each sample.

5. Evaluate the decision rule. Measure the accuracy /error rate in order to improve the choice of features and the overall design of the classifier.

3.4 Performance Evaluation for Predictive Model

Throughout this section the investigator had tacitly assumed that the goal of the performance evaluation was to maximize the success rate of the predictive model for BRHP DSS dataset.

Predictive models are evaluated in terms of correctness, often referred to as performance, and applicability. The performance measures are almost always geared towards the evaluation of an instance of a model type, and are almost always realization method independent. Applicability measures also contain measures that apply to the model type itself, pertaining to the need of models to be evaluated in terms of their context [36].

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

Moreover, the confusion matrix is a useful tool for analyzing how well the researcher's classifier can recognize tuples of different classes. The following procedures and rules were implemented to confirm the model performance evaluation for the results of the predicted model of the under-five mortality in Ethiopia, particularly for BRHP DSS area. Given M classes; a confusion matrix

is a table of at least size M by M. An entry, $CM_{i,j}$ in the first M rows and M columns indicates the number of tuples of class $_i$ that were labeled by the classifier as class $_j$. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry $CM_{1,\ 1}$ to entry $CM_{m,m}$, with the rest of the entries being close to zero [34].

In building a classification model, the confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models [15].

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

Table 3.1: Confusion Matrix

As shown in table 3.1, a confusion matrix table of size two by two, the following measures can be calculated to measure predicted pattern of the under-five model for BRHP DSS dataset's accuracy of the model, True Positive Rate, False Positive Rate, Accuracy, Precision ,Recall, F-measure and ROC Curve.

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

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

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

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

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

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

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

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

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

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

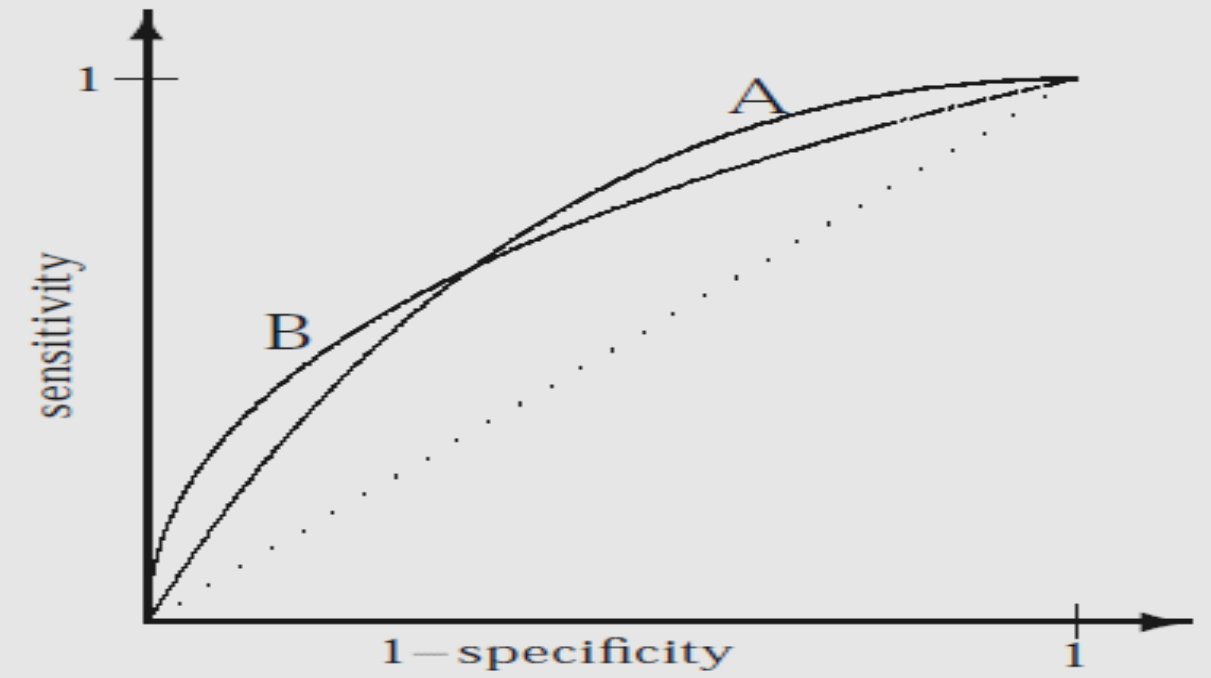
Receiver Operating Characteristics (ROC) Analysis

Another useful method for evaluating classification models is Receiver Operating Characteristics (ROC) analysis. ROC curves are similar to lift charts in that they provide a means of comparison between individual models and determine thresholds which yield a high proportion of positive hits. ROC was originally used in signal detection theory to gauge the true hit versus false alarm ratio when sending signals over a noisy channel [61].

The horizontal axis of an ROC graph measures the false positive rate as a percentage. The vertical axis shows the true positive rate. The top left hand corner is the optimal location in an ROC curve, indicating high TP (true-positive) rate versus low FP (false-positive) rate. The area under the ROC curve (AUC) measures the discriminating ability of a binary classification model. The larger the AUC, the higher the likelihood of an actual positive case will be assigned a higher probability of being positive than an actual negative case. The AUC measure is especially useful for data sets with unbalanced target distribution (one target class dominates the other) [20].

Besides model selection the ROC also helps to determine a threshold value to achieve an acceptable trade-off between hit (true positives) rate and false alarm (false positives) rate. By selecting a point on the curve for a given model a given trade-off is achieved. This threshold can then be used as a post-processing parameter for achieving the desired performance with respect to the error rates.

TP rate



=> FP rate

Figure 3.2: ROC curves. Dotted line has slope 1.

Perfect discrimination gives a curve that is a horizontal line through the point (1, 1), giving an area of 1, while random classification gives a straight curve through the origin with slope of 1, giving an area of 0.5. The binary outcome ROC curve analysis has recently been extended [36] to the case of three classes for which a predictive model returns a probability distribution. The AUC measure has been extended to the volume under the surface (VUS).

Therefore, the researcher tried to implement ROC analysis to evaluate the predicted model constructed by the J48 decision tree algorithm and Naïve Bayes classifier in order to get the best fitted model for the domain area of health sector specially to predict the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS area.

CHAPTER FOUR

BUSINESS UNDERSTANDING AND DATA PREPROCESSING

This chapter provides interesting features for business understanding and data preprocessing of the BRHP DSS dataset emphasizing its ability to accurately measure the under-fives mortality. The current state of BRHP DSS dataset was summarized and the contribution of BRHP DSS dataset to our understanding of under-five mortality in Ethiopia, particularly for BRHP DSS area was discussed briefly.

Under-five mortality has long been used as indicator of the level of socioeconomic development of a nation. Most of the developed countries have registered low levels of under-five mortality rates. The study of under-five children mortality becomes one of the most important researches of the developing countries including Ethiopia. There are two major reasons behind this: (i) high level of infant and child mortality and (ii) its relationship with fertility [60].

Computer assisted information retrieval may help support quality decision making and to avoid human error. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. However, our study is based on computerized health information system which collects, compiles, analyses, interprets, and disseminates health related information for planning and decision making [11].

Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place [56].

To this end a realistic picture of a country's epidemiological profile and the capabilities of its health system are needed before appropriate public health intervention can be developed and implemented. Moreover, intervention developments require an understanding of the determinants of under-five children mortality. The Health Services Extension Program (HSEP) is the main pillar of the child survival strategy for increasing access to promotive, preventive and basic essential curative health services to the majority of the underserved population [1].

Almost 80% of the time and effort in this research project was spent on cleaning and preparing the data for predictive modeling. The BRHP DSS dataset used herein was a single flat file, like SPSS as well as computerized database, such as dBase IV in a fixed-width text format.

4.1 Problem Domain Understanding

Both problem and data understanding are phases in the life cycle of a hybrid DM process. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, was prepared. Finally, project goals were translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

The BRHP centered on Meskan, Mareko and Silti District, Guraghe zone and Silti zone respectively, in the SNNPRS in Ethiopia. The district's population is currently an estimated 260,000, with a density of around 325 people/km^2. The Demographic Surveillance Area (henceforth DSA) covers a sample within the district, following 10 communities initially sampled from the entire district, using a probability proportional-to-size technique. Nine of the 10 sites are rural, and 1 is located in Butajira town [55].

4.1.1 Workflow in the BRHP DSS Area

BRHP DSS is a set of field and computing operations to handle the longitudinal follow-up of well-defined entities or primary subjects (individuals, households, and residential units) and all related demographic and health outcomes within a clearly circumscribed geographic area. Unlike a cohort study, a DSS follows up the entire population of such a geographic area. DSS is an intensive study technique that produces data with substantial advantages over other data [5, 55].

An initial census enumerates and registers the entire population of a well-defined geographic area, the DSA, and after that regular visits are made to each registered location within the DSA in order to record demographic and health-related events that have taken place since the previous visit and to update the status of all entities registered at the location. The DSS study population is typically defined as those people who are resident within the DSA, and it follows that there are only two ways to be admitted to the DSS study population; through birth or in-migration to the DSA. Likewise, there are two ways to exit the DSS study population; through death or out-migration from the DSA. Defined in this way a DSS is similar to a population register; the

primary differences are 1) a DSS usually monitors a comparatively small population intensely and 2) a DSS is an active data collection system that invests considerable effort to track down and visit each member of the study population several times during each year rather than waiting for events and status updates to occur when individuals contact “the system” for some other reason, as is often the case in a population register [55].

The demographic surveillance methodology grew out of the need for accurate information describing the “at risk” (denominator) population living in rural areas in the developing world where vital registration systems either do not exist, or when they exist do not function well enough to provide this information. The primary advantage of most DSSs is that they are the only producers of accurate individual-level, community-based data in the remote rural areas of the Developing World where they are typically (and purposefully) situated [55].

Beyond being the only sources of high quality data in these areas, DSSs produce prospective, fully linked individual, household and community-level data that describe reasonably large whole populations and often include a rich set of prospectively monitored attributes that make possible nuanced longitudinal analyses of mortality at several levels.

By its very nature DSS is longitudinal, and most DSSs collect data over the course of many years thereby describing the history of their study populations. This provides the ability to measure and describe trends in mortality or to control for trends in order to isolate and study other factors contributing to the level of mortality. Because most DSSs visit each registered location several times per year, the temporal resolution and accuracy of the data are both high allowing high resolution trends to be calculated and controlled for; for example seasonal changes in the risk of dying-important in areas where malaria is a significant cause of death [55].

BRHP DSS tracks the presence of individuals in a defined study area. These individuals can enter and leave the study area in a small set of well-defined ways (for example, entering through birth or in-migration and leaving through death or out-migration). The International Network for the continuous Demographic Evaluation of Populations and Their Health in developing countries (henceforth INDEPTH) reference model uses events to record the ways individuals enter (or return to) and leave the study area over time [55].

When a DSS tracks episodes, the concept of the "time resolution" of this tracking is very important. Below a certain time threshold, movements into or out of a particular place are not recorded. DSSs are concerned not only with the physical location or residence of individuals but also with their membership in social groups (such as households) and their relationships with other individuals (such as marital unions or parenthood). Many DSSs also need to reconstruct genealogies and to record isolated events, such as pregnancy outcomes or births and deaths external to the study area [55].

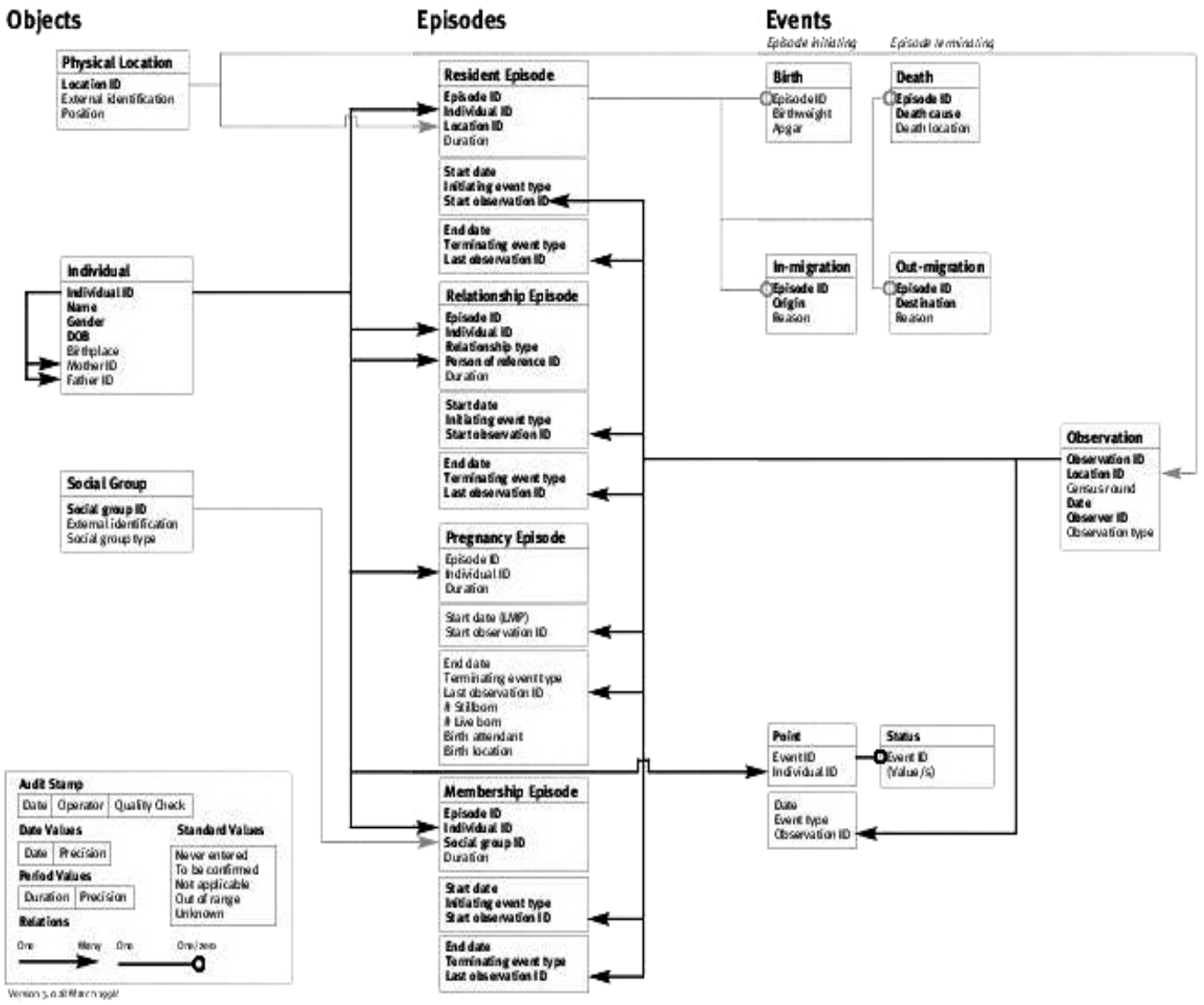


Figure 4.1: Work flow Model of BRHP DSS Area

4.2 Data Understanding

4.2.1 Under-five Mortality Based on BRHP DSS Dataset

During the past 30 years, DSSs have been established in a number of field research sites in various parts of the developing world where routine vital-registration systems were poorly developed or nonexistent. Although these systems may have been developed differently in terms of their initial rationale, they are all required to track a limited and common set of key variables determining population dynamics and demographic trends. DSSs have similar approaches to defining key variables and their relationships and to developing systems for collection, storage, and analysis of these data [55].

Deaths of all registered and eligible individuals are recorded, regardless of the place of death. It may be impossible to record the deaths of previously eligible individuals who then out-migrated. In this case, observation of their survival is censored at the time of migration. Information about the death of visitors to the DSA is sometimes collected, but it is only used in mortality estimates if a de facto population estimate is available for each day [55].

Underreporting of deaths is typically less of a problem than that of births, because a death is widely known and remembered. Exceptions are the deaths of young (and yet unregistered) infants, particularly peri-natal deaths, if cultural beliefs or grief hinders reporting. Some DSSs collect more detailed information about deaths to establish the cause of death, generally through the so-called verbal autopsies (henceforth VAs) [5].

Some consider the under-five mortality as a rate expressing the number of deaths of children <5 years old divided by the number of live births in a year and then multiplied by 1000. Others treat it as a rate, calculating it by dividing the number of deaths of children <5 years old by the total number of person–years of children <5 years old and multiplying by 1000.When under-five mortality is presented as a probability of dying before age of 5 [55].

4.2.2 Data Collection

In BRHP the data was collected by the INDEPTH standards in order to meet the objectives to predict the pattern of under-five mortality in Ethiopia. Any data-collection exercise requires

advance planning and recruitment and training of field staff, such as enumerators and supervisors. It also involves the designing and printing of DSS forms and the preparation of field or training manuals. DSS enumerators are normally recruited from among those local individuals who meet minimum qualifications set for specific projects. Training focuses on proper ways to use DSS forms, conduct interviews, and handle various field forms. Field or interview manuals are used for training and are eventually provided to all field staff as reference materials during data collection [23].

The initial census of the population in the selected villages was done in 1987 to obtain the baseline population and to establish a system of DSS with continuous registration of vital and migratory events at a household level. At that time the total population was 28,780. Any adult member of the household above the age of 15 years was eligible to respond to the monthly household interviews, carried out by a team of secondary school graduate enumerators who were based in the kebeles. Each vital event was registered on a separate form at the household level. Basic demographic, social, housing conditions and health care utilization characteristics were recorded for each household at entry into the DSS and during each re-census process [55].

The longitudinal system of data collection continues then with periodic visits to registered households. Due to circumstances, the first overall update of the 1987 census was not done until 1995, which was, in retrospect, too long interval. A further update round was then conducted in 1999. From the 1987 census until 1999, continuous monitoring was carried out during monthly visits to each household. However, in the light of experience both here and elsewhere, quarterly household visits were phased in during 1999/2000 [23].

4.2.3 Data Source Description

As previously stated, the data was obtained from the BRHP DSS database. The BRHP DSS is primarily a collaborative research project undertaken by the Department of Community Health, Faculty of Medicine, AAU, Ethiopia, and the Division of Epidemiology, Department of Public Health and Clinical Medicine, Umea University, Sweden. The collaboration started as a doctoral-study project [16]. Later, it grew into a departmental collaboration and included the development of the study-base infrastructure and involvement of a multidisciplinary group of researchers. The original DSS population in 1987 was around 28,000 and grew over 10 years to about 37,000

active individuals, with more than 60,000 individuals involved at some time during this first two decades of monitoring [23].

Studies in Butajira have been conducted in a set of nine randomly selected (probability-proportional-to-size technique) rural kebeles (known as "Peasants' Associations") and one urban kebele (the Urban Dwellers' Association). Monthly visits to each household have provided the data. The DSS operates as a dynamic open-cohort system. The individual person–years are aggregated to serve as denominators for calculation of various health and demographic indices. So far, three complete censuses of the population (in 1986, 1995, and 1999) have been done. The extent of similarity between the 1994 national census and the DSS database illustrates the quality of the continuous registration system.

As mentioned earlier in chapter one, the BRHP registers births, deaths, marriages, new households, out- and in-migrations, and internal moves (migration within BRHP DSS kebeles). Household and environmental variables were measured during the censuses. The study base is now well established and is being used for other more focused studies on essential health problems of the country, using qualitative, as well as quantitative, research methods. So far, research on childhood respiratory illnesses, other infectious diseases, reproductive health, and mental health has been conducted using the study-base infrastructure [55]. Mostly the data collected and managed by the BRHP DSS dataset are cleaned; as a result these datasets were less likely to contain missing values [55].

4.2.4 Data Quality Assurance

Data quality assurance mechanisms have been instituted at several points to ensure the integrity of the data. The most critical of these is field supervision. Field supervisors perform the immediate supervision of data collection procedures on a daily basis [23]. Their tasks include checking of each completed data form and visiting randomly selected households each month on a weekly-distributed timetable. The research assistants perform the next level of supervision. They are responsible for the overall supervision of the data flow from the household level to the computer system. Research assistants also perform data checking at the field level in randomly selected households. Researchers also work in the field to provide on-site technical assistance and guidance as well as checking data quality. More recently, with the advent and easy

availability of the Global Positioning System (henceforth GPS), mapping exercises at the household level have been carried out [55].

4.3 Data Preprocessing

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results.

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

Data preprocessing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making [7]. The raw data usually has a great deal of noise. Raw data cannot be used directly for processing, with the machine-learning algorithms. They first need to be preprocessed into machine understandable format. The BRHP's DSS database of under-five children dataset [56] is considered as an example to demonstrate preprocessing.

The initial BRHP's DSS database data which is from 1987-2008 contains the following attributes listed below in table 4.1.

No	Attribute Name	Data Type	Description
1	TTYREF	Char	22-year record reference number
2	PA	Char	PA code for residence episode: 007 Bati,008 Dobena,09B Hobe,09A Mjarda,011 Bido, 04B Dirama, 005 Mmeskan, 06B Wrib, 06A Yeteker, K04 Butajira area 04
3	ENVIR	Char	U = urban, L = lowland, H = highland
4	HOUSENO	Char	house number for residence episode
5	ID	Char	ID number (compulsory) note: IDs beginning "T" are "Old IDs" from the 10 year material, and those beginning "D" are duplicate IDs
6	NAME	Char	Name
7	REL	Char	relationship code: HE = head of household, SP = spouse of head, CH = child of head/spouse, GP = parent of head/spouse, RE = other relative, NR = non-relative, UK = unknown
8	SEX	Char	F = female, M = male (compulsory)
9	MID	Char	mother's id number
10	FID	Char	father's id number
11	MARITAL	Char	marital status during episode: MO = monogamous marriage,PO = polygamous marriage, NM = never married, DI = divorced, SE = separated, WI = widowed, TY = too young, UK = unknown
12	SEREPI	Num	serial no. of individual's episode
13	DBIRTH	Date	date of birth (compulsory)
14	RSTART	Char	reason for episode starting (compulsory) ST = start of surveillance 19870101, BI = birth, IN = in-migration, MO = internal move, LI = changed literacy status, WA = changed source of water, RO = changed

			type of roof, MU = multiple changes without moving, CM = census modification, XX = join between 19951231 & 19960101,CO = continuation from 20050101
15	DSTART	Date	date of episode starting (compulsory)
16	DEND	Date	date of episode ending (compulsory)
17	REND	Char	reason for episode ending (compulsory) EN = end of surveillance 20041231, OU = outmigration, DE = death, MO = internal move, LI = changed literacy status, WA = changed source of water, RO = changed type of roof, MU = multiple changes without moving, XX = join between 19951231 & 19960101, E8 = end of surveillance 20081231
18	DDEATH	Date	date of death (missing: episode not ending in death)
19	TIMEX	Num	days of exposure during episode (>0)
20	CAUSE	Char	cause of death: 1 = measles/chickenpox, 2 = meningitis, 3 = malaria, 4 = tuberculosis, 5 = pneumonia/ARI, 6 = heart disease/sudden death, 7 = tetanus, 8 = jaundice, 9 = diarrhoea/vomiting, A =pregnancy/childbirth/puerperium, B = other cause, C = still-born, D = prematurity, E = malnutrition/kwashiorkor, F = maternal death > 6/52 to 1 yr, G = unknown cause in first month, H = HIV/AIDS, J = accident, Z = unknown, blank = episode not ending in death
21	RELIG	Char	individual's religion: OC = Orthodox Christian, MU = Muslim, CH = Non-orthodox Christian, OT = other, UK = unknown
22	LITER	Char	literacy during episode: LI = literate, RE = reading only, IL = illiterate, TY = too young to be at school, UK = unknown
23	EDUCATION	Char	educational status during episode: NO = no formal education, PR = completed primary school, SE = completed secondary school, TE = further education, UK = unknown

24	SOURCEW	Char	source of water during episode: RI = river, WU = well or spring unprotected, WP = well or spring protected, PI = urban supply (piped), LA = lake or pond, OT = other, UK = unknown
25	ROOF	Char	type of roof during episode: TH = thatched roof, CO = corrugated roof, UK = unknown
26	WINDOWS	Char	windows in the house: YE = yes, NO = none, UK = unknown
27	RADIUS	Num	radius of circular house in metres (missing = 99)
28	ROOMS	Num	number of rooms in house (missing = 99)
29	HOUSEOWN	Char	house ownership: OW = owned, KE = kebele or government, RE = privately rented, OT = other, UK = unknown
30	OXEN	Char	number of oxen owned by family: NO = none, SI = single animal, TW = two or more, UK = unknown
31	TIMAD	Num	number of timad of land owned by family (missing = 99)
32	LATITUDE	Num	latitude of household (missing = 0)
33	LONGITUDE	Num	longitude of household (missing = 0)
34	DISTHOSP	Num	distance to Butajira km (missing = 99)
35	CENSOR	Num	Censored individuals in the BRHP.

Table 4.1: Attributes available in the twenty two years BRHP database

4.3.1 Data Field Selection

Discussing the importance of selecting relevant features (attributes) in any DM task, as cited by Shegaw [11], Liu and Motoda [53] wrote that" the abundance of potential features constitute a serious obstacle to the efficiency of most learning algorithms. Popular methods such as k-nearest neighbor, C4.5, and back propagation are slowed down by the presence of many features, especially if most of these features are redundant and irrelevant to the learning task."

One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models. A very common error, for example, is to use as a predictor variable data that can only be known if you know the value of the response variable.

While in principle some DM algorithms will automatically ignore irrelevant variables and properly account for related (covariant) columns, in practice it is wise to avoid depending solely on the tool. Often your knowledge of the problem domain as well as intensive literature reading can let you make many of these selections correctly [31].

Some of the data or attributes in the initial dataset was not pertinent to the DM goal and were ignored. Of the variables given in the table 4.1 TTYREF, HOUSENO, ID, MID, FID, NAME, SEREPI,RSTART,DSTART,DEND,REND,DBIRTH,MARITAL,CAUSE,LITER,DDEATH,EDUCATION,LONGITUDE,LATITUDE and CENSOR were ignored as having no DM value based on the discussion with domain experts as well as reading literatures from BRHP DSS site and AAU.

In the BRHP DSS database the above listed attributes were not contribute any information towards the machine intelligence in determining whether the under-five children has dead or alive. So, those columns were removed from all the cases within the database.

4.3.2 Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

The first step in data cleaning as a process is discrepancy detection. Discrepancies can be caused by several factors, including poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors (e.g., respondents not wanting to divulge information about themselves), and data decay (e.g., outdated addresses).Discrepancies may also arise from inconsistent data representations and the inconsistent use of codes. Errors in instrumentation devices that record data, and system errors, are another source of discrepancies. Errors can also occur when the data are (inadequately) used for purposes other than originally intended [7].

Moreover, the two-step process of discrepancy detection and data transformation (to correct discrepancies) iterates. This process, however, is error-prone and time-consuming. Some transformations may introduce more discrepancies. Some nested discrepancies may only be detected after others have been fixed. Another approach to increased interactivity in data cleaning is the development of declarative languages for the specification of data transformation operators [7].From the above fact data cleaning task or process is essential. Hence, the researcher tried to carry out different data cleaning tasks in the following sub sections.

4.3.2.1 Handling Missing Values

Missing data is a problem that continues to plague data analysis methods. Even as our analysis methods gain sophistication, the researcher has to continue to encounter missing values in fields, especially in databases with a large number of fields. The absence of information is rarely beneficial. All things being equal, more data is almost always better. Therefore, the researcher considered carefully about how to handle the thorny issue of missing data [54].

Having efficient methods to fill up missing values extends the applicability in terms of accuracy for many DM methods. The accuracy of the tool is increased and with a larger training set better rules and decision trees can be developed which contributes towards better classification of the data to predict the pattern of under-five mortality in Ethiopia, particularly in BRHP DSS area.

After ignoring attributes that have no DM value, the remaining attributes were checked for missing values, inconsistencies and other interpretable observations. The data collected had a small number of variables/attributes with missing values.

A common method of handling missing values is simply to omit from the analysis the records or fields with missing values. However, this may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting records with missing values would lead to a biased subset of the data. Further, it seems like a waste to omit the information in all the other fields, just because one field value is missing. Replace the missing value with the field mean (for numerical variables) or the mode (for categorical variables) [54]. Therefore, in this research study the investigator tried to handle the missing values by replacing missing value with the field

mean, since they are numerical attributes. Table 4.2 summarizes attributes and percentage (%) of missing values associated with each other.

S. No	Attribute Name	Total numbers of Missing Values	Percentage of Missing Values	Mean Value of Missing Values
1	RADIUS	6,855	59.1	3.27
2	ROOMS	387	3.3	1.37
3	TIMAD	2,965	25.6	2.48

Table 4.2: List of variables with their missing values

As a result, the missing values of the dataset were handled in accordance with the above suggestion. The missing value of RADIUS, ROOMS and TIMAD attributes were filled by their mean values since they are numeric value type.

4.3.3 Data Transformation and Reduction

In data transformation, the data were transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

- **Smoothing:** this works to remove noise from the data. Such techniques include binning, regression, and clustering.
- **Generalization of the data:** where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.
- **Normalization:** where the attribute data are scaled so as to fall within a small specified range, such as -1:0 to 1:0, or 0:0 to 1:0. Attribute construction (or feature construction): where new attributes are constructed and added from the given set of attributes to help the mining process.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same)

analytical results [7]. Data transformation is necessary for two purposes to fix problems with the data such as missing values and categorical variables that take on too many values, and to bring information to the surface by creating new variables to represent trends and other ratios and combinations.

In addition, the following data transformation and reformatting operation had employed in order to create new attributes from the existing ones and to reformat the original values of some attributes in the dataset selected for analysis.

A. **Creating the Class attribute:** it is not included in the original dataset instead DDEATH (Date of Death) as an attribute. This derived attribute is a dependent variable can help to classify individuals in to different groups. This classification would help to predict the likelihood that a given individual would dead or alive. Therefore, the class attribute has derived from the DDEATH (Date of Death).

B. **Creating the Age attribute:** in the original database AGE is not included as an attribute in the BRHP DSS database, instead the researcher used DBIRTH as an attribute. However, using AGE as an input (independent) variable can help to categorize the individuals in to different age groups. This categorization would help to identify mortality patterns in individuals with different age groups. Therefore, the age attribute has derived from the DBIRTH (Date of Birth) and also the age attribute has divided in to two sub-categories i.e. infant and child, since a combination of infant and child has known as under-five.

C. **Creating the IN (In-migration) attribute:** it is appeared in the initial dataset as the independent attribute instead the researcher derived from RSTART (Reason to Start) variable. Therefore, the In-migration attribute derived from RSTART and IN had two values, which have either yes or no.

D. **Creating the OU (Out-migration) attribute:** it has appeared in the initial dataset as the independent attribute instead the researcher derived from REND (Reason to End) variable. Therefore, the Out-migration attribute derived from REND also OU had two values which have either yes or no.

Therefore, the researcher discretized some attributes by converting numeric attributes to nominal: specify which attributes, number of bins, whether to optimize the number of bins, and output binary attributes in order to get the best-fitted model. Table 4.3 provides summary of the original attributes and derived attributes with their values.

No	Original Attributes	Derived Attributes	Values
1	DBIRTH	AGE	String
2	RSTART	IN	Char
3	REND	OU	Char
4	DDEATH	CLASS	Char

Table 4.3: Summary of derived attributes with their values

The final selected dataset with their description are summarized in table 4.4.

No	Attribute Name	Data Type	Description
1	PA	Char	PA code for residence episode: 007 Bati,008 Dobena,09B Hobe,09A Mjarda,011 Bido, 04B Dirama, 005 Mmeskan, 06B Wrib, 06A Yeteker, K04 Butajira area 04
2	ENVIR	Char	U = urban, L = lowland, H = highland
3	REL	Char	relationship code: HE = head of household, SP = spouse of head, CH = child of head/spouse, GP = parent of head/spouse, RE = other relative, NR = non-relative, UK = unknown
4	SEX	Char	F = female, M = male (compulsory)
5	TIMEX	Num	days of exposure during episode (>0)
6	RELIG	Char	individual's religion: OC = Orthodox Christian, MU = Muslim, CH = Non-orthodox Christian, OT = other, UK = unknown
7	SOURCEW	Char	source of water during episode:

			RI = river, WU = well or spring unprotected, WP = well or spring protected, PI = urban supply (piped), LA = lake or pond, OT = other, UK = unknown
8	ROOF	Char	type of roof during episode: TH = thatched roof, CO = corrugated roof, UK = unknown
9	WINDOWS	Char	windows in the house: YE = yes, NO = none, UK = unknown
10	RADIUS	Num	radius of circular house in metres (missing = 99)
11	ROOMS	Num	number of rooms in house (missing = 99)
12	HOUSEOWN	Char	house ownership: OW = owned, KE = kebele or government, RE = privately rented, OT = other, UK = unknown
13	OXEN	Char	number of oxen owned by family: NO = none, SI = single animal, TW = two or more, UK = unknown
14	TIMAD	Num	number of timad of land owned by family (missing = 99)
15	DISTHOSP	Num	distance to Butajira km (missing = 99)
16	AGE	Num	Age
17	IN	Num	In migration
18	OU	Num	Out migration
19	CLASS	Char/String	Classes of the dependent/outcome attribute

Table 4.4: Final selected variables with their description

After the successful preprocessing of the required dataset from the original twenty two years' data, the next important issue considered by the researcher was importing the selected, created and reformatted dataset, which were in SPSS, STATA and Excel document format into WEKA software which has an understandable format to construct a best fitted model to predict the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS area.

4.3.4 Machine Understandable Format in WEKA

Most DM tools can use data in the CSV format for running the machine intelligent algorithms. A common situation is for the data to be stored in a spreadsheet or database. However, WEKA expects it to be in ARFF format, because it is necessary to have type information about each attribute which cannot be automatically deduced from the attribute values. Before you can apply any algorithm to your data, is must be converted to ARFF form. This can be done very easily. Recall that the bulk of an ARFF file consists of a list of all the instances, with the attribute values for each instance being separated by commas (Table 4.5) [52].

Most spreadsheet and database programs allow you to export your data into a file in CSV format-as a list of records where the items are separated by commas. Once this has been done, you need only load the file into a text editor or a word processor; add the dataset's name using the @relation tag, the attribute information using @attribute, and a @data line [52]. In the Table 4.5, the last attribute where the classification of individual is done is made into a categorical format. The last attribute where the classifications attribute 'class' takes string values 'Dead' when death occur and 'Alive' when death is not occur.

@relation 'BRHPDataset'

@attribute PA {007,008,09B,09A,011,04B,005,06B,06A,K04}
@attribute ENVIR {U,L,H}
@attribute REL {HE,SP,CH,GP,RE,NR,UK}
@attribute SEX {M,F}
@attribute TIMEX numeric
@attribute RELIG {OC,MU,CH,OT,UK}
@attribute SOURCEW {RI,WU,WP,PI,LA,OT,UK}
@attribute ROOF {TH,CO,UK}
@attribute WINDOWS {YE,NO,UK}
@attribute RADIUS numeric
@attribute ROOMS numeric
@attribute HOUSEOWN {OW,KE,RE,OT,UK}

@attribute OXEN {NO,SI,TW,UK}

@attribute TIMAD numeric

@attribute DISTHOSP numeric

@attribute AGE numeric

@attribute IN {Yes, No}

@attribute OU {Yes, No}

@attribute CLASS {DEAD, ALIVE}

@data

5,H,CH,F,934,MU,WU,TH,NO,3.27,1,OW,UK,2,2.8,3,No,No,Alive

5,H,CH,F,221,MU,WP,CO,YE,3.27,2,OW,NO,1,3,1,Yes,No,Dead

5,H,UK,F,373,MU,WP,CO,YE,3.27,2,OW,SI,4,3,3,No,Yes,Alive

5,H,CH,M,98,MU,WP,CO,YE,3.27,2,OW,SI,4,3,0,Yes,Yes,Alive

5,H,CH,M,17,MU,RI,TH,NO,3,1,OW,SI,3,3.1,4,No,Yes,Alive

Table 4.5: Sample WEKA System Understandable ARFF Format for BRHP Dataset

In this research work, the performance of the classifier was evaluated by using the most common test option, i.e. cross validation and % split test options. Thus, by invoking these options, training and testing cases were selected from the BRHP DSS database file.

CHAPTER FIVE

EXPERIMENTATION AND ANALYSIS

This chapter is devoted to discuss on the models to be built and experiments carried out together with their analysis. The experiments were run on a larger dataset in order to address the main objectives of the research study with respect to the minimum data set that consists of 19 attributes. This will help in understanding the different stages that were used in various DM algorithms.

In this study an attempt was made to design a model that enables to predict the pattern of under-five mortality in Ethiopia. To this end, classification i.e. J48 decision tree and Naïve Bayes classifiers were selected and experimented on. BRHP DSS database was consulted to extract the dataset required for training and testing the models created by the classifiers. For creating predictive model a total size of 11,600 records were used for training and testing. The validations were done using 10-fold cross validation and 90% split test option.

5.1 Dataset Preparation

The data collected for this research project came from BRHP DSS database was in SPSS format. The dataset initially had 35 attributes and 320,112 records but after preprocessing stage, it was reduced to 19 attributes and 11,600 records for building the predictive model for under-five children. Preprocessing was computed on SPSS software that was extracted from BRHP DSS database. Then after, the preprocessed dataset converted to Comma Separated Values (.csv) and then Attribute Relation File Format (.arff) which was compatible with WEKA software for model building.

A dataset of BRHP was imbalanced if the classification categories are not approximately equally represented [62]. Performance of DM algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large. In the case of BRHP DSS under-five mortality data, the class variable status has a higher imbalance. Therefore, the researcher used Synthetic Minority Oversampling Technique

(henceforth SMOTE) automatic operation by filter where minority classes are over sampled by generating synthetic examples of minority class and adding them to the dataset. This way, the class distribution in the dataset changes and probability of correctly classifying minority class increases [34].

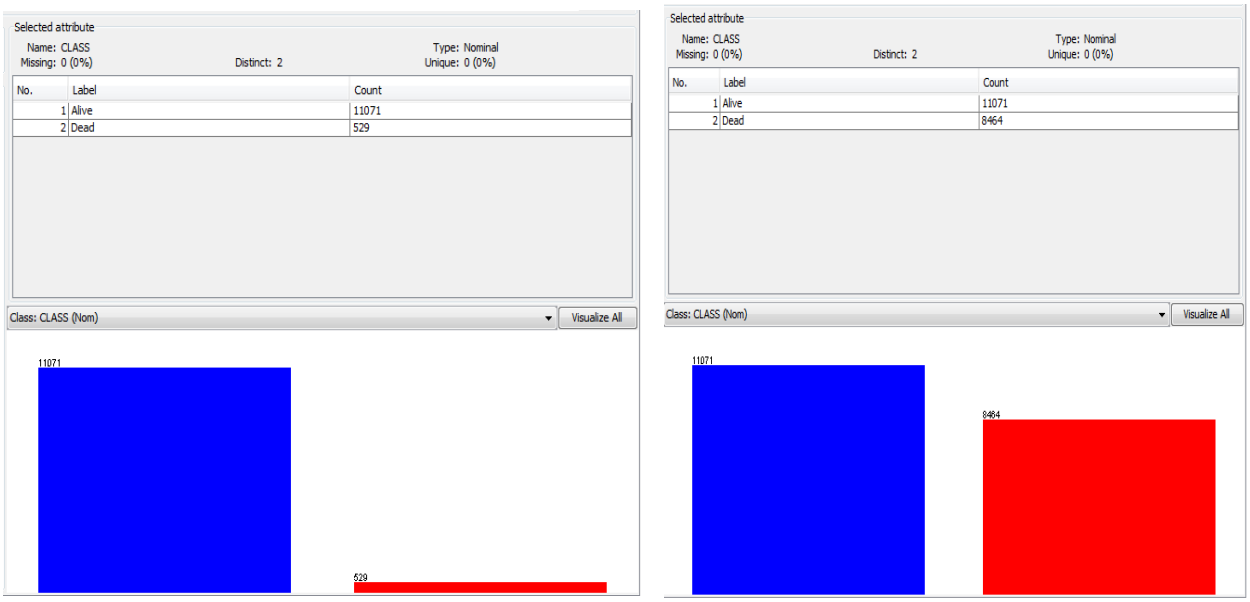


Figure 5.1: Side by side view of the class variable: (a) Original data; (b) Balanced data using SMOTE.

Figure 5.1 shows a side by side review of the class attribute status after SMOTE operation applied to the minority class. Originally there were 11,071 records in the majority class and only 529 records in the minority class but after applying SMOTE the difference between the two classes were reduced only to 2,607 records.

5.2 Model Building

The researcher took 19 attributes: 18 features+1class attribute for building predictive model. The selection of attribute was made using subjective judgment of the investigator, reviewing literature and discussion with public health professionals from BRHP and AAU. To build the predictive model, the .arff and/or .csv format of the selected dataset was given to WEKA: they are J48 decision tree algorithms and Naïve Bayes classifiers.

5.2.1 Building Classification Model using WEKA Software

Classification is learning a function that maps (classifies) a data item in to one of several predefined classes. In relation to this, Chapman, P. et al. [64] noted that "when learning classification rules, the system had to find the rules that predict the class-label, which is the dependent or predicted attribute's value, from the independent or predicting attributes' value". Accordingly, it was the concern of this phase to generate classification rules that were assigning the correct class label to previously unseen and unlabeled children's.

In this phase, the main issue to be discussed was to build a classification model. Due to the fact that well-grown decision trees were as comparably useful as other classifiers, the type of classification model selected to be built was decision tree [34]. The other reason for selecting decision tree model is, compared to other classification model types like Neural Network classification, decision tree has a significant advantage because it can be built manually-and so, is easily explained [18]. Apart from this, decision tree operations are completely interactive, iterative and they benefit from powerful visualization features.

To carry out this phase, the well known software, WEKA 3-6-2 had been attempted to use. Therefore, further discussion in this investigation was done by making use of the models found from the selected WEKA's algorithms runs [34].

5.2.1.1 J48 Decision Tree Model Building Using WEKA Software

The J48 decision tree C4.5 algorithm builds decision trees from a set of predefined training dataset using the concept of information entropy and attribute ordering. It uses the fact that each attribute of the data was used to make a decision by splitting the data into smaller subsets.

As decision tree is a classifier, any previously unseen record with the required degree of attributes was fed into the tree. At each node, it will be sent either left or right to some test.

Finally, it will reach a leaf node and be given the label associated with that leaf. At this junction, the researcher interested to generating rules of assigning the under-five children's of BRHP DSS members to the class they belong.

The classification C4.5 algorithms were implemented in WEKA 3-6-2 to build a classification model in such a way that testing the model would be possible after training it. For most of the experiments carried out in this phase, the experiments were 10-fold cross validation and Percentage (%) split test options were used, the total record was partitioned in to two, the training and test datasets. These two datasets were found from the final data set by using a stratified sampling technique where the different classes, found in the classification, were considered as strata for 10-fold cross validation.

The reason behind applying stratified sampling technique was, partitioning the total record where the contribution of each class in the resulting datasets was proportional. To avoid the problem of over-fitting the researcher tried to experiment by pruning the tree, 10% of the total record was selected as a test sub-data set and the remaining 90% as a training sub-data set.

5.2.1.1.1 The WEKA Decision Tree Experiment and Analysis

WEKA 3-6-2 supports many types of classification algorithms. Among the classification algorithms that WEKA 3-6-2 supports, the J48 C4.5 algorithm was used with different input parameters as well as different types of related classifiers. J48 algorithm is WEKA's implementation of the C4.5 decision tree learner. The corresponding algorithm used to extract rules from the decision trees is J48 or PART.

By making use of WEKA 3-6-2 a total of 16 experiments were carried out, where 4 of the experiments were for constructing decision trees with 10-fold cross validation, 6 were different values of percentage split test, 4 were for constructing decision trees with 90% split test and the remaining 2 experiments were for Naïve Bayes classifier with or without supervised discretization respectively. In relation to this, J48 was the algorithm used to construct the decision trees in the 14 experiments. The extraction of the corresponding rules, in the remaining 14 experiments, from the decision trees was managed using J48 or PART.

To display the run parameters and the outputs of the respective experiments, three tables (Table 5.1, Table 5.2 and Table 5.3) were used. As displayed in all tables, the different experiments were carried out by using all the 19 attributes of the records with different schemes were applied in the experiment and two different test modes (ways of feeding records to the algorithms).

Analysis of the J48 decision tree predictive model were made in terms of detailed accuracy, precision, recall, F-measure and ROC curve of the classifier based on a confusion matrix of each predictive model resulted of different classes (Alive and Dead classes in this research thesis).

The experiment number from 1 up to 4 were applied on the 10-fold cross validation test mode, 6 experiments (i.e. from 5 up to 10) were relied on different value of percentage split test mode and the remaining 4 experiments (i.e. from 11 up to 14) were relied on 90% split test mode . The experiments for J48 decision tree classification models are listed under beneath:

Experiment 1: Unpruned J48 decision tree with default confidence factor (i.e. 0.25) and with 10-fold cross validation test mode.

Experiment 2: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and with 10-fold cross validation test mode.

Experiment 3: Unpruned J48 decision tree with confidence factor 0.15 and with 10-fold cross validation test mode.

Experiment 4: Pruned J48 decision tree with confidence factor 0.15 and with 10-fold cross validation test mode.

Experiment 5: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 66% split test mode.

Experiment 6: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 70% split test mode.

Experiment 7: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 80% split test mode.

Experiment 8: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 85% split test mode.

Experiment 9: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 90% split test mode.

Experiment 10: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 95% split test mode.

Experiment 11: Unpruned J48 decision tree with default confidence factor (i.e. 0.25) and

90% split test mode.

Experiment 12: Pruned J48 decision tree with default confidence factor (i.e. 0.25) and 90% split test mode.

Experiment 13: Unpruned J48 decision tree with confidence factor 0.15 and 90% split test mode.

Experiment 14: Pruned J48 decision tree with confidence factor 0.15 and 90% split test mode.

These experiments were analyzed to compare them in terms of different performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC/AUC curve. The models were also compared with regard to the pattern or KD of the predictive model.

Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. The sensitivity is the measure of the ability of a prediction model to select instances of a certain class from a data set. The specificity corresponds to the true negative rate which is commonly used in two class problems [10].

As an outcome of different combinations of these scheme and J48 algorithm was used in the experiments. These combinations were: J48 decision learner algorithm and J48 and/or PART decision rule extractor algorithm.

The two J48 decision tree test modes were:

- **T1**-Inputting all the records with a 10-fold cross-validation test mode, and
- **T2**-Inputting all the records Percentage(%) split test which is train a model and then supply the unseen remaining part of the record for testing the performance of the model.

Furthermore, the result obtained from these experiments was summarized in table 5.1 with respective performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC/AUC curve.

S. No	Comparing Parameters	Experiments No			
		1	2	3	4
1.	Testing Mode	T1	T1	T1	T1
2.	Pruning	No	Yes	No	Yes
3.	Confidence Factor	0.25	0.25	0.15	0.15
4.	Size of Tree	1199	327	1199	267
5.	No. of Leaves	881	215	881	175
6.	Time Taken (Sec.)	1.14	1.42	0.91	1.05
7.	Precision	0.972	0.97	0.972	0.97
8.	F-Measure	0.973	0.975	0.973	0.975
9.	Sensitivity	0.974	0.981	0.974	0.979
10.	Specificity	0.964	0.96	0.964	0.96
11.	AUC	0.976	0.981	0.976	0.982
12.	Accuracy (%)	96.93	97.19	96.93	97.1

Table 5.1: Input parameters and the resulting J48 Decision Trees' with 10-fold CV test mode.

As can be observed from Table 5.1, the number of leaves and the corresponding sizes of the trees constructed from experiments 2 and 4 are less than those found form the other experiments. As the investigator can be inferred from the above table that experiment 2 has higher accuracy and other performance measurement than the rest experiments except specificity measurement. Therefore, experiment 2 outperforms than the other experiments in performance for BRHP DSS datasets. The data collected, preprocessed and analyzed using classification (J48 decision tree) was presented in the below table 5.2. The researcher tried to classify with different values of percentage (%) split test parameters of trained and tested data to look the performance of the system. The following are some samples of the experiments

Experiment	Split Test Mode in %	Accuracy in %
5.	66	96.91
6.	70	96.83
7.	80	96.83
8.	85	97.07
9.	***90***	***97.49***
10.	95	97.13

Table 5.2: Input parameters and resulting J48 DT with different percentage split test mode.

As it can be observed from the above table, the 90% split test of data for training is better than the other percentages split test options. The selected percentage has 97.49% correctly classified instances. Percentage split test parameter of 95% training set has also 97.13 % correctly classified instances, but with relatively low precision, recall, F-measure and ROC curve.

Moreover, the J48 decision tree model produced was from the table 5.2 has experiment 2 which has 90% split test mode which is train a model and then supply the unseen remaining part of the record for testing the performance of the model and its accuracy level was 97.49 %. The 90% test option mode shows a better performance than others experiments.

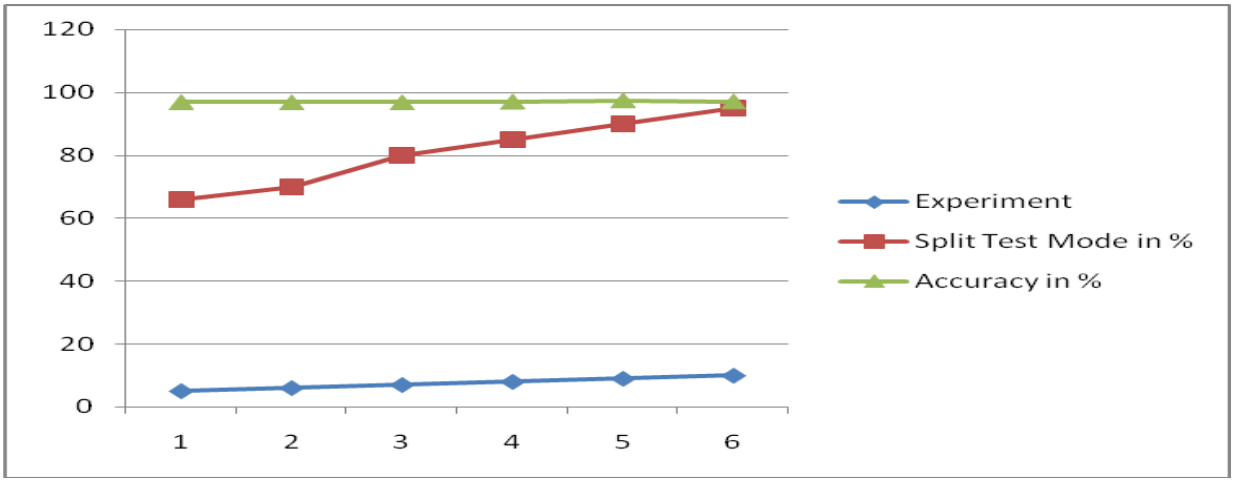


Figure 5.2: Line Graph of J48 Decision Trees' with different percentage split test mode options.

In Figure 5.2 the researcher showed that the line graph of the accuracy obtained for the different tools. The highest accuracy is obtained by the 90% split test method. Thus, it is considered also

the base case. All the other tools tested have performed lesser than the 90% split test method. The accuracy on an average for the test of the tool is 97.49%.

The percentage split test option was used to partition the dataset into training and testing data and this parameter was set to 90, 90% for training and 10% for testing. The result of this learning scheme was summarized and presented in table 5.3. Moreover, the result obtained from these experiments is summarized in table 5.3 with respective performance matrices values, accuracies, size of trees, no. of leaves, time taken in sec. in the execution, and ROC/AUC curve.

S. No	Comparing Parameters	Experiments No			
		11	12	13	14
1.	Testing Mode	T2	T2	T2	T2
2.	Pruning	No	Yes	No	Yes
3.	Confidence Factor	0.25	0.25	0.15	0.15
4.	Size of Tree	1199	327	1199	267
5.	No. of Leaves	881	215	881	175
6.	Time Taken (Sec.)	0.92	1.08	0.94	1.08
7.	Precision	0.975	0.974	0.975	0.975
8.	F-Measure	0.976	0.978	0.976	0.977
9.	Sensitivity	0.977	0.982	0.977	0.98
10.	Specificity	0.966	0.965	0.966	0.966
11.	AUC	0.98	0.988	0.98	0.986
12.	Accuracy (%)	97.24	97.49	97.24	97.39

Table 5.3: J48 Decision Trees' with 90-percentage split test mode parameters.

As can be observed from this Table 5.3, experiments 12 and 14 have comparatively better than the other experiments with extracted accuracies and rules. This is because, the researcher used to build J48 decision tree with default confidence factor (i.e. 0.25) and 90% split test mode and also the pruned parameter of the classifier. This result portrays that due to the adjustment of some of the parameters, the size of tree reduced to 327 and the number of leaves has become 215.

The model has accuracy of 97.19% using 10-fold cross-validation and 97.49% accuracy using 90% split test options. Moreover, the model has a true positive rate of 98.1% and true negative rate of 96% for 10-fold cross-validation and also a true positive rate of 98.2% and true negative rate of 96.5% for 90% split test options.

The best J48 decision tree model of the classification generated from experiment 12 of the 90% split test mode. The model shows a better performance evaluation than other models. The 90% split test model also scored a better performance than 10-fold cross-validation. Therefore, the test options mode used to build the decision tree for experiment 12 with 90% split test mode options which is J48 pruned decision tree with default confidence factor (i.e. 0.25), are statistically significant in splitting the decision tree. Furthermore, suggestions gathered from the domain experts from AAU and literatures indicated that these attributes have a great role in the prediction tasks.

Though, experiments 12 and 14 were examined, respectively, with experiments 11 and 13, in the table above, they were simply done by feeding the algorithms with 90% split test mode. Therefore, from Table 5.3, since experiment 12 was carried out to construct the required decision tree with a 90% split test mode and had a reasonably good accuracy, it was selected. In addition to this, experiment 12, which is the corresponding rule extraction experiment from the J48 decision tree constructed, was selected. In this regard, generally, the reasons of selecting experiments 12 from all the experiments carried out could be mentioned follows:

- The number of records considered is relatively large.
- The number of attributes selected and used is effective.
- The number of leaves and size of the tree in experiment 12 are manageable; and the number of rules extracted in experiment 12 is reasonable.
- The test mode, which is the 90% split test mode, used in the experiment 12 is acceptable.
- The accuracy of the resulting model is comparatively better than others.

As a result, the full outputs of the selected working J48 decision tree in experiment 12 and the corresponding rules extracted in experiment 12 was annexed for reference.

Confusion Matrix for J48 Decision Tree Model

A confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models. Moreover, the overall predictive accuracy on unseen instances it is often helpful to see a breakdown of the classifier's performance.

The confusion matrix for J48 decision tree presented underneath in table 5.4 depicts that out of the total records provided to the program 1,105 (98.2%) and 799(96.5%) records were classified correctly in the class of Alive and Dead respectively. On the other hand, 20 (1.8%) records were incorrectly classified as "Dead" while actually they were supposed to be in the "Alive" class and 29 (3.5%) records were classified incorrectly as Alive while actually they are in the Dead class. This portrays that from the total records 1,904 (97.49%) records were classified correctly while the remaining 49(2.51 %) records were classified incorrectly. Hence, this indicated that records whose class is "Alive" were classified with a minimum error as compared with the records in the class "Dead".

Actual Class	Predicted Class		Total
	Alive	Dead	
Alive	1,105	20	1,125
Dead	29	799	828
Total	***1,134***	***819***	***1,953***

Table 5.4: Confusion Matrix for J48 decision tree model

Consecutively, to see how well the predictive model can recognize "Alive" tuples (the positive records) and how well the predictive model which has the classifier can recognize "Dead" tuples (the negative records) which have sensitivity and specificity measures can be used. Sensitivity is also known as the true positive cases in the test data with predicted probabilities greater than or equal to the probability threshold (correctly predicted), while specificity is the true negatives rate: Negative cases in the test data with predicted probabilities strictly less than the probability threshold (correctly predicted). Furthermore, the classifier has 98.2% sensitivity and 96.5%

specificity; which discloses that the J48 decision tree classifier has an acceptable capability of recognizing the true class value.

ROC Analysis for J48 Decision Tree Model

ROC curves are similar to lift charts in that they provide a means of comparison between individual models and determine thresholds which yield a high proportion of positive hits. In the below figure the horizontal axis of an ROC graph measures the false positive rate as a percentage. The vertical axis shows the true positive rate. The top left hand corner is the optimal location in an ROC curve, indicating high TP (true-positive) rate versus low FP (false-positive) rate.

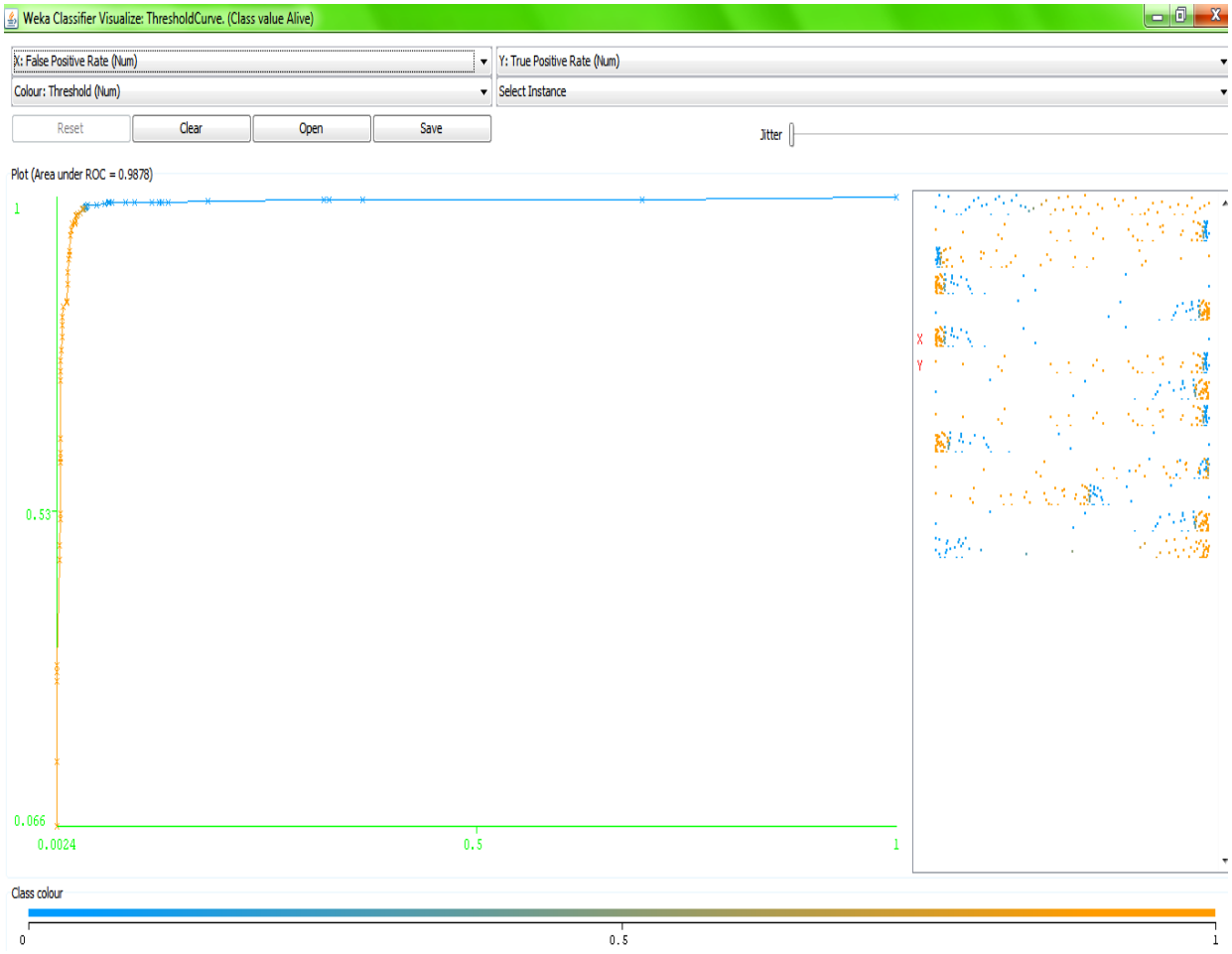


Figure 5.3: ROC curve of the J48 decision tree model

The true positive (TP) rate and false positive (FP) rate values of different classifiers on the same test set are often represented diagrammatically by a ROC Graph. The abbreviation ROC analysis

stands for ‘Receiver Operating Characteristics Graph’, which reflects its original uses in signal processing applications. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. Class value Alive gives the ROC accuracy of 98.78%. The larger the area under the ROC curve (AUC), the higher the likelihood that an actual positive case and also the better the model. It will be assigned a higher probability of being positive than an actual negative case. The AUC for the model is 0.9878 which is closer to 1, that is AUC measure is especially useful for data sets with unbalanced target distribution (one target class dominates the other).

Figure 5.4 shows that, the tree view of the predictive model built J48 algorithm with 90% split test mode options which is J48 pruned decision tree with default confidence factor (i.e. 0.25) using 19 attributes. For clear understanding of the tree, the run information for the predictive model is annexed at Annex III.

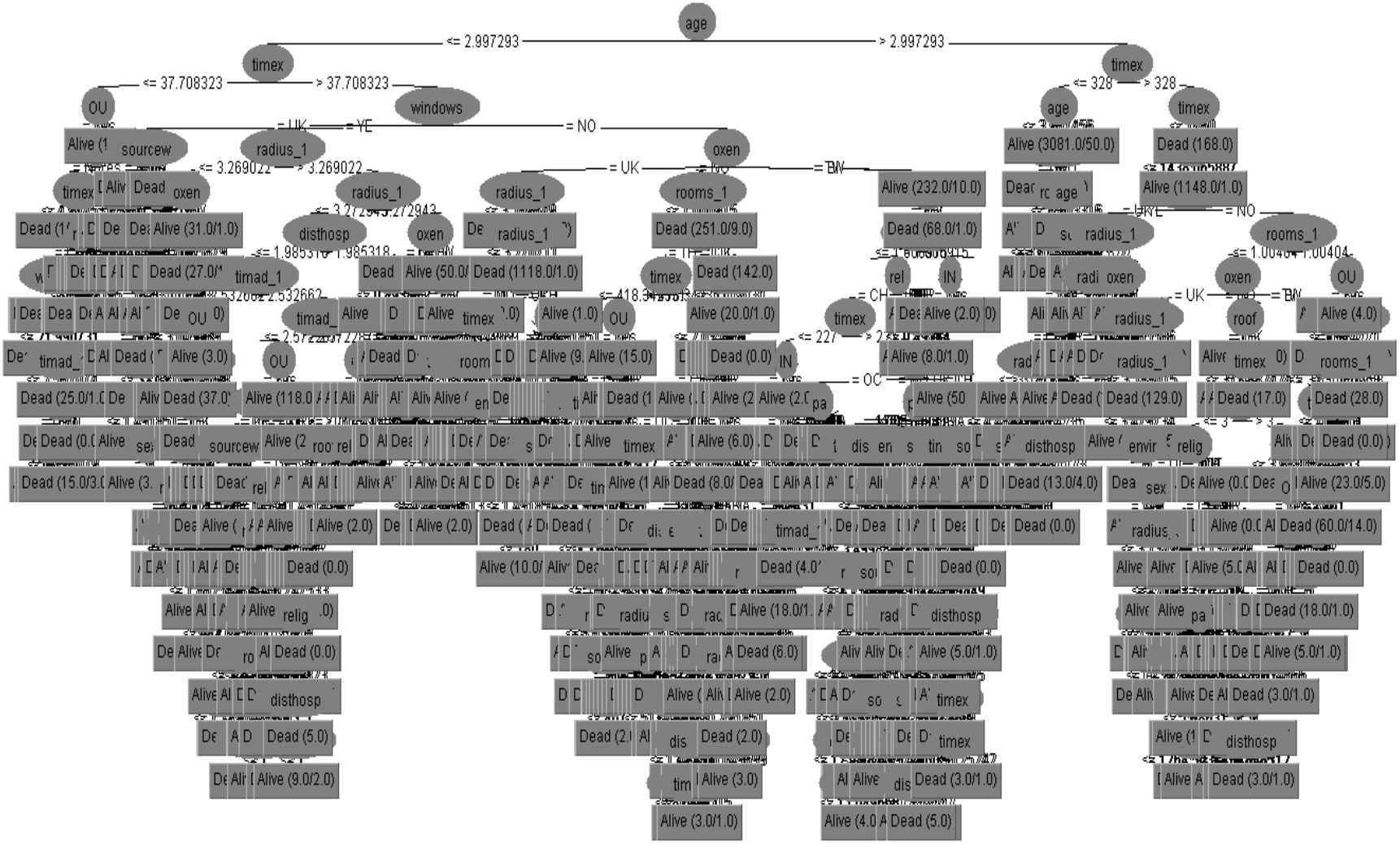


Figure 5.4: Tree view of the Predictive Model Using J48 Algorithm

5.2.1.2 Naïve Bayes Classifier Model Building using WEKA Software

It is method of classification that does not use rules, a decision tree or any other explicit representation of the classifier. Rather, it uses the branch of Mathematics known as probability theory to find the most likely of the possible classifications. The Naïve Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which the researcher used to calculate the probability of each of the possible classifications in turn. Having done this the researcher chooses the classification with the largest value. Taking into account the nature of the underlying probability model, the Naïve Bayes classifier can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect [5, 6].

In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data. However, various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domains.

The Naïve Bayes [7] classifier provides a simple approach, with clear semantics, to representing and learning probabilitistic knowledge. It is termed naïve because is relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it posits that no hidden or latent attributes influence the prediction process. Two experiments were done with the Naïve Bayes algorithm using different combination and inputting all records with an inputting 90% split test with and without supervised discretization parameter, which is train a model and then supply the unseen remaining part of the record for testing the performance of the model. Table 5.5 summarizes the results with respective performance matrices values, accuracies, discretization parameter, time taken in sec. in the execution, precision, F-Measure, sensitivity, specificity and ROC curve.

Experiment 15: T3-Naïve Bayes classifiers with 90-percentage split test mode without supervised discretization.

Experiment 16: T3-Naïve Bayes classifiers with 90-percentage split test mode with supervised discretization.

S. No	Comparing Parameters	Experiments No	
		15	16
1.	Testing Mode	T3	T3
2.	Discretization.	No	Yes
3.	Time Taken (Sec.)	0.09	0.41
4.	Precision	0.848	0.947
5.	F-Measure	0.783	0.972
6.	Sensitivity	0.726	0.998
7.	Specificity	0.824	0.924
8.	AUC	0.885	0.985
9.	Accuracy (%)	76.75	96.67

Table 5.5: Summary of Naïve Bayes Experiment Results

As can be observed from the above table 5.5, the model scored in experiment no.15 was 76.75% accuracy, 72.6% true positive rate and 82.4% true negative rate using 90% split test mode without supervised discretization and the model scored in experiments no. 16 was 96.67% accuracy, 99.8% true positive rate and 92.4% true negative rate using 90% split test mode with supervised discretization. Hence, experiment 16 with all attributes and 90% split test option with supervised discretization were chosen to have a better performance evaluation for Naïve Bayes classifier.

Confusion Matrix for Naïve Bayes Classifiers

A confusion matrix, sometimes called a classification matrix, is used to assess the prediction accuracy of a model. It measures whether a model is confused or not; that is, whether the model is making mistakes in its predictions. Various classification rules were used in creating a confusion matrix. The classification rules that

incorporate prior probabilities, posterior probabilities and misclassification costs are based on Bayesian statistical decision theory. Bayesian theory essentially revises prior probabilities based on additional available information [52].

It is the result of a test task for classification models. Moreover, the overall predictive accuracy on unseen instances it is often helpful to see a breakdown of the classifier's performance. Table 5.6 shows the Naïve Bayes form of a confusion matrix that is used for calculating goodness of fit and goodness of prediction errors. True Positive is defined as the case in which the test result and gold standard (truth) are both positive; False Positive is the case in which the test result is positive but the gold standard is negative; True Negative is the case where both are negative; and False Negative is the case where the test result is negative but the gold standard is positive.

Actual Class	Predicted Class		**Total**
	Alive	Dead	
Alive	1,123	2	**1,125**
Dead	63	765	**828**
Total	**1,186**	**767**	***1,953***

Table 5.6: Confusion Matrix for Naïve Bayes model

The confusion matrix for Naïve Bayes classifier presented in the above table 5.6 depicts that out of the total records provided to the program 1,123 (99.8%) and 765(92.4%) records were classified correctly in the class of Alive and Dead respectively. On the other hand, 2 (0.2%) records were incorrectly classified as 'Dead' while actually they were supposed to be in the 'Alive' class and 63 (7.6%) records were classified incorrectly as 'Alive' while actually they are in the 'Dead' class. This portrays that from the total records 1,888 (96.67%) records were classified correctly while the remaining 65(3.33%) records were classified incorrectly. Hence, this indicated that records whose class is 'Alive' were classified with a minimum error as compared with the records in the class 'Dead'.

Furthermore, the classifier has 99.8% sensitivity and 92.4% specificity; which reveal that the Naïve Bayes classifier has an immense capability of recognizing the true class value.

ROC Analysis for Naïve Bayes Classifiers

ROC analysis is performed by drawing curves in two dimensional spaces, with axes defined by the True Positive rate and False Positive rate, or equivalently, by using terms of sensitivity and specificity. That is, the y-axis represents Sensitivity = True Positive rate, while the x-axis represents 1−Specificity = False Positive rate.

The AUC for the under-five mortality records generated from the Naïve Bayes Classifier is presented in the below figure 5.5. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. Class value Alive gives the ROC accuracy of 98.55%.

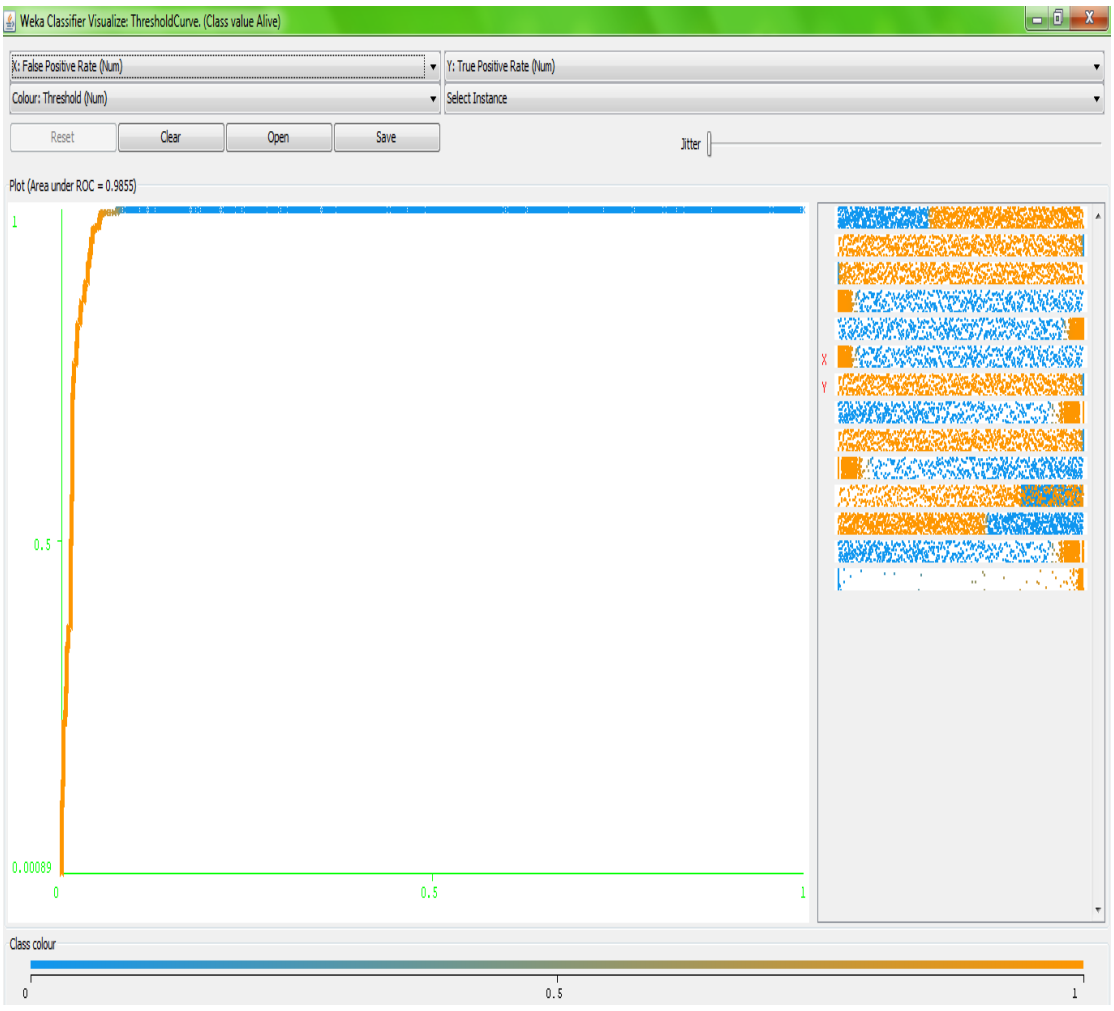


Figure 5.5: ROC curve from the Naïve Bayes Classifier

In the above figure the horizontal axis of an ROC graph measures the false positive rate as a percentage. The vertical axis shows the true positive rate. The top left hand

corner is the optimal location in an ROC curve, indicating high TP (true-positive) rate versus low FP (false-positive) rate. ROC plots allow for visual comparison of several models (classifiers). For each model, the researcher calculated its sensitivity and specificity, and draws it as a point on the ROC graph.

In Figure 5.5, the area under the diagonal curve is 0.9855. Thus, the researcher interested in choosing a model/classifier that has maximum area under its corresponding ROC curve: the larger the area, the better performing the model/classifier is. There exists a measure similar to the AUC for assessing the goodness of a model/classifier, known as the Gini coefficient, which is defined as twice the area between the diagonal and the ROC curve; the two measures, however, are equivalent, since Gini+1 = 2 AUC.

5.3 Analysis and Discussion of the Classification Model

In this research project work, several experiments had been carried out with two classification algorithms, i.e. J48 decision tree algorithm and Naïve Bayes classifier to build a predictive model that predicts the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS area. From the experiments all attributes were identified to make sound rule and better accuracy. Both classifiers algorithms were compared due to inputting 90% split test which is train a model and then supply the unseen remaining part of the record for testing the performance of the model.

From the confusion matrix to analyze the performance criterion for the J48 decision tree algorithm and Naïve Bayes classifiers are summarized in table 5.7 in the predicting of the under-five mortality, accuracy, precision (for multiclass dataset), Area under the ROC, Time taken for execution, sensitivity and specificity have been computed to give a deeper insight of the automatic diagnosis [5].

And also, a comparison of the performance evaluation in the table 5.7 between J48 decision tree algorithm and Naïve Bayes classifier are illustrated in figure 5.6.

Performance Testing	J48 Decision Tree	Naïve Bayes Classifier
Accuracy (%)	97.49	96.67
Precision (%)	97.4	94.7
Time taken for execution (sec.)	1.08	0.41
Sensitivity (%)	98.2	99.8
Specificity (%)	96.5	92.4
F-Measure (%)	97.8	97.2
AUC (%)	98.8	98.5

Table 5.7: Performance comparison of J48 Decision Tree and Naïve Bayes classifier with 90% split test mode.

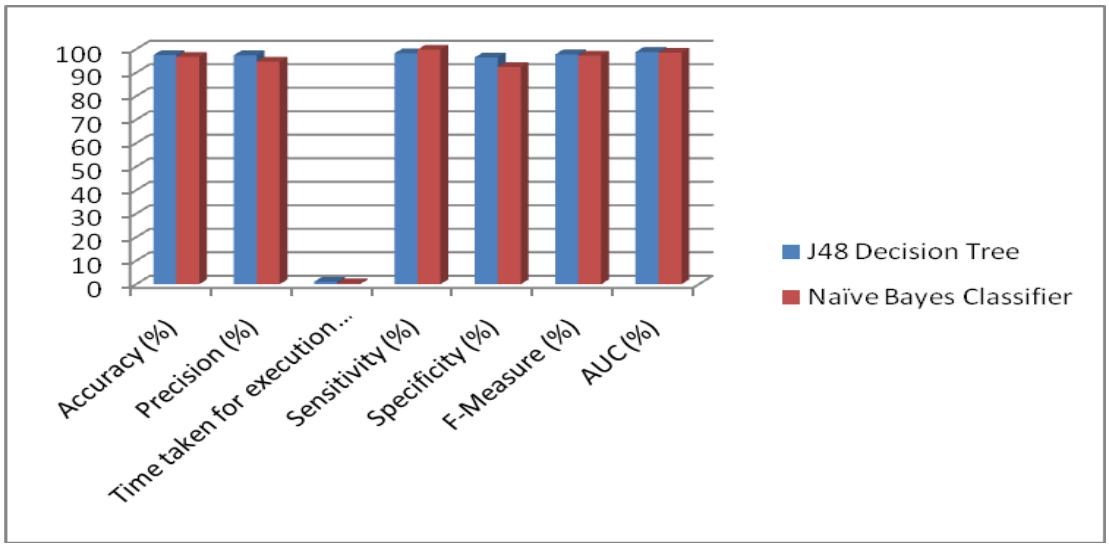


Figure 5.6: Bar Graph Visualization of Performance comparison of J48 Decision Tree and Naïve Bayes classifier with 90% split test mode.

One of the purposes of this study was to compare the J48 decision tree algorithm and Naïve Bayes classifier DM model and to select the one, which performs the best. Accordingly, each experiment carried out in this research had employed both J48 decision tree and Naïve Bayes classifier. In all experiments the same datasets were used. The output of these experiments indicated that the classification task of records using the under-five mortality dataset from BRHP DSS area.

In Figure 5.6 the researcher visualized the bar graph of the performance evaluation obtained for the different tools. The highest accuracy is found by the J48 decision tree method. Thus, it is considered also the base case. All the J48 decision tree algorithm tools tested have performed much better than the Naïve Bayes classifier method.

The result scores of the Naïve Bayes classifier for time taken to execute the model have better than the J48 decision tree model. However, the overall result scores of the J48 decision tree model higher than that of the Naïve Bayes classifier model.

In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, Time taken for execution, AUC, sensitivity and specificity). The results were achieved using inputting 90% split test which is train a model and then supply the unseen remaining part of the record for testing the performance of the model. In comparison to the above studies, the researcher found that the predictive model achieved a classification accuracy of 0.9749 with a sensitivity of 0.982, time taken to execute in sec. has 1.08, AUC is 0.988, and a specificity of 0.965.

5.4 Classifier Error

A classifier is a model of data used for a classification purpose: given a new input, it assigns that input to one of the classes it was designed/ trained to recognize.

When the investigator talk about a model, there is always a model error associated with it. Model error is calculated as the difference between the observed/true value and the model output value, and is expressed either as an absolute or squared error between the observed and model output values. The researcher generated a model of the dataset that fitted to the dataset. However, in addition to fitting the model to the dataset, the investigator interested in using the model for prediction. Thus, once the investigator had generated several models and had selected the “best” one, the investigator need to validate it, not only for its goodness of fit (fit error), but also for its goodness of prediction (prediction error). Data-reuse, or re-sampling, methods (simple split, cross-validation, bootstrap) are very popular in evaluating supervised learning models [66].

Sometimes, the predicted and actual value may differ in predicting a record to a certain class label. This shows that the record that is labeled by an expert to one class may be labeled by the classier to other class. This kind of features often reduces the performances of the system.

The model built using J48 decision tree algorithm with 90% split test option mode used a total of 1,953 records for testing the model performance. 1,105 records and 799 records were correctly classified as “Alive” and “Dead” respectively. The classifier incorrectly classified 20 records as “Dead” and 29 as “Alive”. The below table shows Sample of instances that show the actual and predicted class difference.

S. NO.	AGE	TIMEX	OU	IN	WINDOWS	RADIUS	TIMAD	DISTHOSP	ENVIR	CLASS	
										ACTUAL	PREDICTED
1.	3	934	No	No	No	3.27	2	2.8	H	Alive	Alive
2.	4	58	No	No	No	3	2	3	H	Alive	Dead
3.	2	89	Yes	No	Yes	3.27	3	7.6	L	Alive	Dead
4.	3	1080	No	No	Yes	3.27	2.48	0.7	U	Alive	Dead
5.	4	995	No	No	Yes	3.27	2	15.3	L	Alive	Dead
6.	2	1	No	No	No	2	3	3.1	H	Dead	Alive
7.	1	1	No	Yes	No	3.27	2.48	3.8	H	Dead	Alive
8.	3	19	No	No	Yes	3	4	10.4	L	Dead	Alive
9.	3	1101	No	No	Yes	3.27	2.48	0.9	U	Dead	Alive
10.	3	4	No	Yes	No	4	4	4.8	H	Dead	Dead

Table 5.8: Sample of records that show the actual class and predicted class variation

As shown from the above table, the experts classified the pattern of under-five children survival status whether Alive or Dead based on BRHP DSS database. Then the classifier wrongly classified the status of the under-five children. As it is observed from the evaluation results on test data, the total error rate of this classifier was 2.509% where it wrongly classified 29 ‘Dead’ cases as ‘Alive’ and 20 ‘Alive’ cases as ‘Dead’.

5.5 Generating Rules from J48 Decision Tree

From the decision tree developed in the aforementioned experiments, it is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node [65]. This produces rules that are unambiguous in that it doesn't matter in what order they are executed. Decision tree and decision rule solutions offer a level of interpretability that is unique to symbolic models. The solutions may be directly inspected to understand the decision surfaces that exist in the data [48].

When the researcher compared the performance measure as well as the results obtained from both decision tree algorithms i.e. J48 and PART models are nearly equal performance. In terms of accuracy, execution time, AUC, sensitivity and specificity; J48 is slightly better than PART. Therefore, the researcher selected the J48 decision tree algorithm for generating better rules. The following are some of the rules extracted from the J48 decision tree are listed below and some of the rules suppose to be interesting and are selected by domain experts as well as from the literatures, are presented as follows:

Rule: 1

If AGE is '<=2.99' and TIMEX is with 4.94-37.71 and OU (Out-migration) is No and IN (In-migration) is No and REL is CH (Child of Head/Spouse) and WINDOWS are YE (Yes) then the class of the under-five children is likely to be ***'Dead'*** *(156.0/10.0).*

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

=>156/(156+10)= 156/166= 0.94= 94%

Rule: 2

If AGE is '<=2.99' and TIMEX is '<=37.71' and OU (Out-migration) is No and IN (In-migration) is Yes and ROOF is TH (Thatched) and OXEN is SI (Single) then the class of the under-five children is likely to be ***'Dead'*** *(4.0/1.0).*

Rule: 3

If AGE is '<=2.99' and TIMEX is with 37.71-187.45 and WINDOWS are YE (Yes) and RADIUS is '<=3.27' and OXEN is SI (Single) and ROOMS are '<= 1' and ROOF is TH (Thatched) then the class of the under-five children is likely to be ***'Dead'*** *(17.0/5.0).*

Rule: 4

If AGE is '<=2.99' and TIMEX is with 37.71-187.45 and WINDOWS are YE (Yes) and RADIUS is '<=3.27' and OXEN is TW (Two) then the class of the under-five children is likely to be ***'Alive'*** *(31.0/1.0).*

Rule: 5

If AGE is '<=2.99' and TIMEX is '> 37.71' and WINDOWS are YE (Yes) and RADIUS is '<=3.27' and DISTHOSP is '<=1.99'and TIMAD is '<= 2.53' and OU (Out-migration) is No and ROOF is TH (Thatched) then the class of the under-five children is likely to be ***'Dead'*** *(170.0/5.0).*

Rule: 6

If AGE is '<=2.99' and TIMEX is with 37.71-418.34 and WINDOWS are NO (None) and OXEN is NO (None) and ROOMS are '<=1' and ROOF is TH (Thatched) and OU (Out-migration) is No and REL is CH (Child of Head/Spouse) and TIMAD is '<=1' then the class of the under-five children is likely to be ***'Alive'*** *(14.0/2.0).*

Rule: 7

If AGE is '<=2.99' and TIMEX is with 37.71-418.34 and WINDOWS are NO (None) and OXEN is NO (None) and ROOMS are '<=1' and ROOF is TH (Thatched) and OU (Out-migration) is No and REL is CH (Child of Head/Spouse) and TIMAD is with 0.11-1 and RADIUS is '<=3.27' then the class of the under-five children is likely to be ***'Dead'*** *(118.0/30.0).*

Rule: 8

If AGE is '<=2.99' and TIMEX is with 37.71-418.34 and WINDOWS are NO (None) and OXEN is NO (None) and ROOMS are '<=1' and ROOF is TH (Thatched) and OU (Out-migration) is No and REL is CH (Child of Head/Spouse) and TIMAD is with 0.11-1 and RADIUS is '>3.27' and ENVIR is H (Highland) then the class of the under-five children is likely to be ***'Dead'*** *(22.0/1.0).*

Rule: 9

If AGE is '<=2.99' and TIMEX is with 37.71-418.34 and WINDOWS are NO (None) and OXEN is NO (None) and ROOMS are '<=1' and ROOF is TH (Thatched) and OU (Out-migration) is No and REL is CH (Child of Head/Spouse) and TIMAD is with 0.11-1 and RADIUS is '>3.27' and ENVIR is L (Lowland) then the class of the under-five children is likely to be ***'Alive'*** *(3.0/1.0)*.

Rule: 10

AGE is with 1.01-2.99 and TIMEX is '> 37.71' and WINDOWS are NO (None) and OXEN is NO (None) and ROOMS are '<=1' and ROOF is TH (Thatched) and OU (Out-migration) is No and REL is CH (Child of Head/Spouse) and TIMAD is '<=2.57' and RADIUS is '>3.63' then the class of the under-five children is likely to be ***'Dead'*** *(6.0/2.0)*.

Rule: 11

If AGE is '<=2.99' and TIMEX is '> 37.71' and WINDOWS are NO (None) and OXEN is TW (Two) then the class of the under-five children is likely to be ***'Alive'*** *(232.0/10.0)*.

Rule: 12

If AGE is '<=2.99' and TIMEX is '> 37.71' and WINDOWS are NO (None) and OXEN is SI (Single) and ROOF is TH (Thatched) and ROOMS are '>1' and IN (In-migration) is No then the class of the under-five children is likely to be ***'Dead'*** *(126.0/7.0)*.

Rule: 13

AGE is with 2.99-3.97 and TIMEX is '<=328' and WINDOWS are YE (Yes) and ROOF is TH (Thatched) then the class of the under-five children is likely to be ***'Alive'*** *(68.0/2.0)*.

Rule: 14

AGE is with 2.99-3.97 and TIMEX is '<=328' and WINDOWS are NO (None) and ROOF is TH (Thatched) and OU (Out-migration) is No and REL is RE (Other Relative) and IN (In-migration) is Yes then the class of the under-five children is likely to be ***'Alive'*** *(11.0/1.0)*.

Rule: 15

If AGE is '>2.99' and TIMEX is with 328-1460 and WINDOWS are YE (Yes) and RADIUS is '>3.27' and OXEN is SI (Single) and SOURCEW is WU (Well or Spring Unprotected) then the class of the under-five children is likely to be ***'Dead'*** *(9.0/2.0)*.

The above rules indicate how a given record could be classified based on some attribute values to construct rules and provided the class predicted by the rule. Hence, having these rules, instances were classified into the predefined classes. In fact, in classifying under-five children records into the predefined classes, from the all attributes such as RELIG, PA, HOUSEOWN and SEX were not occurred in the above generated rules which have a base for the classification tasks. The numerical value, which written at the end of the predicted class in bracket, indicates the level of assurance of the independent attribute for the status or the predicted class.

5.6 Discussion of Results on Classification Models from Generated Rules

The following discussion or explanation on the generated rules was made with the domain experts from BRHP of AAU as well as from the literatures.

Some of the rules presents known pattern as the domain experts opinion (the rules generated was agreed with domain experts view as well as from literatures). If under-five children in BRHP whose age is less than or equal to 2.99 and whose days of exposure during episode (TIMEX) is between 37.71 & 187.45 and whose windows in the house are yes and whose circular house radius is less than or equal to 3.27 and also whose number of oxen owned by family is two then that leads to the 'Alive' class.

To add another known rule, if under-five children in BRHP whose age is less than or equal to 2.99 and whose days of exposure during episode (TIMEX) is greater than

37.71 and whose windows in the house are yes and whose circular house radius is less than or equal to 3.27 and the distance to Butajira hospital is less than or equal to 1.99 km and also whose type of roof is thatched then that leads to the ‘Dead’ class.

It also known, if under-five children in BRHP whose age is less than or equal to 2.99 and whose days of exposure during episode (TIMEX) is between 37.71 & 418.34 and whose windows in the house are none and whose number of oxen owned by family is none and whose number of rooms in house are less than or equal to 1 and whose type of roof during episode is thatched and the out-migration level is no and whose relationship is child of head or spouse and also whose number of timad of land owned by family is less than or equal to 1 then that leads to the ‘Alive’ class.

On the other hand, an interesting rule shows, if under-five children in BRHP whose age is greater than to 2.99 and whose days of exposure during episode (TIMEX) is between 328 & 1460 and whose windows in the house are yes and whose circular house radius is greater than to 3.27 and whose number of oxen owned by family is single and also the source of water during episode is well or spring unprotected then that leads to the ‘Dead’ class.

Another interesting rule shows, if under-five children in BRHP whose age is less than or equal to 2.99 and whose days of exposure during episode (TIMEX) is between 37.71 & 418.34 and whose windows in the house are none and whose number of oxen owned by family is none and whose number of rooms in house are less than or equal to 1 and whose type of roof is thatched and the out-migration level is no and whose relationship is child of head or spouse and also whose number of timad of land owned by family is between 0.11 & 1 and whose circular house radius is greater than to 3.27 and also the environment is highland then that leads to the ‘Dead’ class.

Generally, it is possible to say that some of the rules obtained from the predictive model provide a pattern or knowledge and have got meaningful contribution for exploring the outcomes of the classes’ in the BRHP and these findings also got acceptance by the domain experts from BRHP of AAU. Consequently, to evaluate the significance of the above selected rules from the generated model and the attributes used to construct those rules, the relationship of the attributes with the predicted class

as well as the model predicted by rules was evaluated based up on suggestions given by domain experts and reports from the literatures.

As it is observed from rules 1 up to 15 above, under-five mortality is related with age of the child, days of exposure during episode (timex), out-migration, in-migration, relationship, availability of windows in the house, type of roof during episode, number of oxen owned by family, radius of circular house in metres, number of timad of land owned by family, distance to Butajira hospital, number of rooms in house, environment and source of water during episode.

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

As it is presented in rule 8 and rule 9, environmental variations between highland, lowland and urban communities in BRHP DSS exposed marked differences in under-five mortality. As per the discussions made with public health experts from AAU and pediatricians from Tikur Anbessa Hospital, the researcher confirmed that these variations in under-five mortality among highland, lowland and urban communities seems appropriate since, there is an epidemic of malaria and meningitis in rural lowlands than in highlands as well as urban communities. The under-five mortality rate for children who live in rural areas is 30% higher than that for children who live in urban areas.

From rules 1 up to 15, except rule 2 windows was identified as a determinant factor for the survival of the under-five child. The classifiers constructed by using rule set have also revealed that lack of windows (i.e. poor housing conditions and crowding in the house) as a determinant factor for under-five children mortality in the BRHP DSS area. Particularly, Infants (whose age is below 1 year) are more vulnerable for mortality due to lack of windows in the house. Similar findings have been observed in other studies conducted in the study area. For example, a study conducted by Desta [16] indicated that out of 128 deceased infants, 101 lived in houses without a window. This study has also demonstrated that for infants, a fivefold ARI (Acute Respiratory Infection) mortality risk is associated with lack of a window.

From the above all rules, except rule 1, 4 and 15 type of roof was also observed that also identified as the major determinants of under-five mortality in rural communities of the BRHP DSS area. According to the WHO literatures for dwelling units, about 85% of the "Tukuls" in a rural area are crowded and far behind to satisfy the physiological needs of a resident. Nearly 95% of housing units had only one room. The high magnitude of overcrowding in rural housing units indicates the poor living and sanitation conditions [21]. With respect to wealth and mortality, the relationship is not consistent, although children born to mothers in the higher timad of land & oxen wealth clearly are at much lower risk of dying than children born to mothers in the lower timad as well as oxen. According to NCSSE, 70 % of the deaths before age five, the cause is a disease that would be preventable in a high-income country.

As it can be seen from Rule 5 up to 10, number of timad of land owned by family were identified as a determinant factor for the survival of the under-five children. With respect to wealth and mortality, the relationship is not consistent, although children born to mothers in the highest timad of land wealth quintile clearly are at much lower risk of dying than children born to mothers in the other quintiles.

From rules 3 up to 15, radius of circular house was identified as a determinant factor for the survival of the under-five children. Therefore, based on those rules, they can be deduced that children whose parents' radius of circular house values are lesser they are more vulnerable to mortality than those whose parents' are radius of circular house values are higher. The importance of this variable to determine the risk of child mortality has been reported in various studies. According Desta [16] pneumonia is the leading cause of child deaths accounting for 40% of deaths in this age group. Availability of Windows a in the house revealed a statistically significant association with under-five children survival. If no ventilation in the house, the there is a high possibilities of under-five mortality due to Pneumonia.

An interesting finding identified from the above rules, particularly from Rules 5 was distance to Butajira hospital is the relationship of under-five children mortality. As it is observed in this rule, the result of distance, large proportions of under-five children family were forced to spend many hours to get to the nearest health facility and this is particularly the case in children whose parents are from BRHP DSS area. Therefore,

the longer the distance from the hospital, the higher likely the under-five mortality in BRHP DSS area.

The source of water for the child's parents was also identified as a determinant factor for survival of the under-five child. Particularly, as it is observed from rule 15 if the source of water for the under-five child's family is well or spring unprotected, they are more vulnerable to mortality than those whose source of water is well or spring protected. There is ample evidence that access to adequate and safe water and sanitation can influence under-five children mortality.If lack of adequate and safe water and sanitation can influence under-five children mortality due to diarrhoea.

Parental factors affected the infants relatively more than they did the children, especially with regard to ARI mortality. This was also noted with "absence of window", a proxy measure for evaluating the type of housing. In terms of etiological fractions, a greater number of under-five children deaths could be ascribed to parental than environmental conditions, with relatively more infants being affected than children. As it can be observed from the previous discussions, the ruleset option had derived interesting and useful rules that can be applied to predict the pattern of under-five survival probability of infants and children in the BRHP DSS area.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

Machine intelligence algorithms are improving as the number of DM tools, techniques and algorithms increase. Healthcare data is a good test bed for DM. A great deal of data in health care is still being gathered and organized using pen and paper. Indeed, the data contains and reflects activities and facts about the organization. But, the data's hidden value, the potential to predict health trends, has largely gone unexploited. The increase in data volume causes great difficulties in extracting useful information and knowledge for decision support. It is to bridge this gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as DM or KDD has emerged in recent years.

The application of DM technology has increasingly become very popular and proved to be relevant for many sectors such as healthcare sectors. Particularly, in the public health, DM technology has been applied for predicting the pattern of under-five mortality for effective and efficient predictive model, determinants and patterns that contributes to the occurrence of under-five children mortality.

This research has tried to assess the application of DM technology to predict the pattern of under-five mortality in Ethiopia, for developing a classification model. Such a classification model could enable the public health department of BRHP as well as for the governmental and non-governmental organizations to implement predictive model in Ethiopia.

This investigation, conducted according to the hybrid KDP model, was carried-out in six major parts namely: business understanding, data understanding, data preparation, model building, evaluation, and use of the discovered knowledge. However, since a DM task is an iterative process, these steps were not followed strictly.

A data set with 11,600 total BRHP members' records was used to develop a classification model. Since, this research was intended to fill a gap left by a related research; some valuable experiences of the previous research were used. However, though the previous research's objective was to identify children at risk for certain ailments by using the same data, the classification indexes were not documented along with each individual record.

In the classification phase, the J48 decision tree algorithm and Naïve Bayes classifier, which is WEKA's implementation of the C4.5 algorithms, were used. From the study, AGE, TIMEX, WINDOWS, TIMAD, OXEN, SOURCEW, ROOF, ROOM, DISTHOSP & ENVIR attributes were the significant DM and Public Health values.

In order to select a classification model that can classify the BRHP members, the models were built by employing the J48 decision tree algorithms. In the decision tree selection process, more emphasis was given to important attributes to be used, the number of records considered and the size of the tree and the corresponding number of rules extracted from the tree. Though, the number of attributes selected and used for the under-five mortality, both in the previous and this investigations, were 34, only 19 of these attributes were found to be useful for the current classification tasks.

From the experiments done using a WEKA version 3-6-2, it was observed that, for a given number of attributes, as the number of records used to develop a decision tree increases the corresponding number of rules generated will possibly increase. Due to this observation, not to get a minimum number of rules, among all the models developed for comparison, the models developed from the 11,600 records with a 90% split test option and attribute selection were given due attention. Accordingly, the better J48 decision tree with the corresponding extracted rules was selected as a working model to classify members into their corresponding classes. As a result, the classification accuracy of the selected J48 decision tree seems convincing than the Naïve Bayes classifier. That is, among the 11,600 data inputted to the model learner with a 90 % split test mode, 97.49%, which is 1,904 records were correctly classified.

The suggestions and opinions given by domain experts in the entire investigation were observed and found to be very important in the model development process, particularly, in the classification phase.

The overall predictive model building process made by employing the J48 decision tree algorithm and Naïve Bayes classifier demonstrated that DM is a method that should be considered to predict the pattern of under-five mortality in Ethiopia, particularly for BRHP DSS area.

6.2 Recommendations

This investigation has been conducted mainly for an academic purpose. However, it revealed the potential applicability of DM technology to classify children whose age category has under-five in the BRHP DSS dataset. Moreover, it is the researcher's belief that the contribution of this research work could be a good experience for a competitive study in public health as well as information science sectors of under-five children in the future.

Apart from this, it is the researcher's faith that the findings of the research would encourage public health sector to work on the application of DM technology to appreciate and employ under-five children's survivals, and as a result gain a competitive advantage based on demographic, socio-economical, parental, environmental, and epidemiological factors alone.

Therefore, the researcher strongly recommends the following:

- In this research encouraging results were obtained, further investigation should be done by integrating the numerous under-fives children survival data sources.
- Programs should expand on health education would increase access of people to information, and improve the access of safe water, poverty reduction strategies and environmental protection.
- There is a need to develop an operational application prototype named under-fives children survival classification system.
- Further extensive experiments should be required by using large amounts of dataset and applying different classification techniques.

- There is a need different DM research investigations based on clinical datasets from different health facilities.
- There should be different DM researches can be undertaken by comparing the DHS dataset with DSS dataset.
- Further study is recommended to the problem domain specifically and under-fives children in general that apply those unused DM models, tools and algorithms.

REFERENCE

1. Federal Ministry of Health (FMOH). (2005). National Strategy for Child Survival in Ethiopia, Addis Ababa, Ethiopia.
2. United Nations Development Programme (UNDP). (2007). Measuring Human Development: A primer, New York: UNDP, USA.
3. Central Statistical Agency (CSA) [Ethiopia] and ORC Macro. (2011). Ethiopia Demographic and Health Survey 2011: Preliminary Report. Addis Ababa, Ethiopia and Calverton, Maryland, USA: CSA and ORC Macro.
4. Central Statistical Agency (CSA) [Ethiopia] and ORC Macro. (2006). Ethiopia Demographic and Health Survey 2005. Addis Ababa, Ethiopia and Calverton, Maryland, USA: CSA and ORC Macro.
5. Berhane, Y. and Byass, P. (2003). Butajira DSS Ethiopia, Department of Community Health, Faculty of Medicine AAU and Department of Public Health and Clinical Medicine Umea University, INDEPTH Monograph: Volume I Part C.
6. Hian, C. K. and Gerald, T. (2005).Data Mining Applications in Healthcare, London, UK.
7. Han, J. and Kamber, M. (2006). Data Mining: concepts and Techniques. 2nd ed. Morgan kufman Publishers, San Francisco, USA.
8. Piatetsky-Shapiro, G. and Frawley, W. (1991). Knowledge Discovery in Databases, AAAI/ MIT Press, MA, USA.
9. Ministry of Finance and Economic Development (MoFED). (2010). Millennium Development Goals Report 2010. Addis Ababa, Ethiopia.
10. Emamu, A. (2011).Mining Emergency Medical Data: The Case of Tikur Anbessa Specialized Hospital. MSc. Thesis, Addis Ababa University, Ethiopia.
11. Shegaw, A. (2002). Application of data mining technology to predict child mortality patterns: the case BRHP, MSc. Thesis, Addis Ababa University, Ethiopia.
12. David, H. et. al. (2001).Principles of Data Mining. MIT Press, London, UK.

13. Abera, K. (2006). Retrospective cohort study in the determinants of child mortality in BRHP and DSS. MSc. Thesis, Addis Ababa University, Ethiopia.
14. Cios, K. et. al. (2000). A knowledge discovery approach to diagnosing myocardial perfusion: IEEE Engineering in Medicine and Biology Magazine, New York, USA.
15. Amir, F. and Shahram, J. (2011). An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network. International Journal of Advanced Science and Technology. Vol. 34, Shiraz University, Shiraz, Iran.
16. Desta, S. (1994). Epidemiology for public Health research and action in a Developing society: the BRHP in Ethiopia. EJHD, 8 (Special Issue).
17. Taddesse, B. (2011). Mining Vital Statistics Data: The case of BRHP. MSc. Thesis, Addis Ababa University, Ethiopia.
18. Theeuwen, M. et. al. (2001). Neural Network analysis to predict Treatment outcome in patients with ovarian cancer. Available URL: http://www.mbfys.kun.nl/mbfys/people/bert
19. Julio, P. and Adem, K. (2009). Data Mining and Knowledge Discovery in Real Life Applications, I-Tech pub. , Vienna, Austria.
20. Emmanuel, N. O. (2007). Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. Turks & Caicos Islands Community College.
21. World Health Organization (WHO). (2008). World health statistics 2008. Geneva, Switzerland.
22. Fayyad, U. et. al. (1996a). The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 39, 11, 27-34. New York, USA.
23. Berhane, Y. et. al. (2004). Impact of child mortality and fertility preferences on fertility status in rural Ethiopia. East Africa Medical Journal 81(6):300-6. Addis Ababa, Ethiopia.
24. Berry, M. A. and Linoff, G. (1997). Data Mining Techniques: for Marketing, Sales, and Customer support. John Willy& Sons Inc, New York, USA.

25. Patricia, C. (2010).Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks, University of Louisville, USA.
26. Immon, W. (1996). Building the Data Warehouse. New York: John Wiley and Sons, Inc., USA.
27. Mitchell, M. (1997). Machine Learning. New York: The McGraw Hill, USA.
28. Max, B. (2007). Principles of Data Mining .Springer, New York, USA.
29. Velickov, S. and Solomatine, D. (2000). Predictive Data Mining: Practical Example, Moscow, Russia.
30. Sumathi, S. and Sivanadam, S.N. (2006). Introduction to Data Mining and its Application .Berlin: Springer, Germany.
31. Two Crows Corporation. (2005). Introduction to Data Mining and Knowledge Discovery.3rd ed., New York, USA.
32. Glasnik, M. (2008). Data Mining Usage in Healthcare Management: Literature Survey and Decision Tree Application –Original Article, Volume 5, Number 1, Cavtat, Croatia.
33. Chang, C. L. (2007). A study of applying DM to early intervention for developmentally-delayed children. Expert Systems with Applications. Huwei, Taiwan.
34. Selam, A. (2011). Predicting the Occurrence of Measles Outbreak in Ethiopia Using DM Technology. MSc. Thesis, Addis Ababa University, Ethiopia.
35. Julio, P. and Adem, K. (2007).Data Mining and Knowledge Discovery in Real Life Applications, Printed in Croatia.
36. Vinterbo, S. A. (1999). Predictive Models in Medicine: Some Methods for Construction and Adaptation. Norwegian University of Science and Technology, Oslo, Norway.
37. Alexander, K. S. et. al. (2001). Hybrid Decision Tree Learners with Alternative Leaf Classifiers: An Empirical Study. Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria.
38. Aha, D. et. al. (1991). Instance-Based Learning Algorithms. Machine Learning 6(1), Washington DC, USA.

39. Cover, T. M. and Hart, R. E. (1997).Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theoo'IT-13(I), London, UK.
40. John, G. H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. Proc. of 11th Conference on Uncertainty in Artificial Intelligence 338-345. Montreal, Canada.
41. Langley, P. et. al. (1992). An Analysis of Bayesian Classifiers. In Proceedings of the Tenth National Conference on Artificial Intelligence 223-228. AAAI Press/MIT Press, Cambridge/Menlo Park. Montreal, Canada.
42. Quinlan, J. R. (1986). Induction of Decision Trees: Machine Learning 1(1):81-106. Edinburgh University Press.
43. Quinlan, J. R. (1993). C4.5; Programs for Machine Learning. Morgan Kaufmann, Los Altos/Palo Alto/San Francisco, USA. His web site is URL: http://www.rulequest.com
44. Witten, I. et. al.(1999). Data Mining. Morgan Kaufmann, Los Altos/Palo Alto/San Francisco, USA.
45. Everitt, B. S. and Hand, D. J. (1981). Finite Mixture Distributions. London: Chapman and Hall, UK.
46. Odei, S. D. (2006). An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain. MSc. Thesis, Bournemouth University.
47. Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules, in Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile.
48. Apte, C. and Weiss, S. M. (1997). Data Mining with Decision Trees and Decision Rules. T. J. Watson Research Center IBM Research Division Yorktown Heights, NY 10598, New York, USA.
49. Dunham, M. (2003). Data Mining: Introductory and Advanced Topics, Prentice Hall, Upper Saddle River, New Jersey, USA.
50. Lewis, R. (2002). An introduction to classification and regression tree (CART) analysis, in Annual Meeting of the Society for Academic Emergency Medicine. San Francisco, CA, USA.

51. Michael, W. B. and Murray, B. (2006). Lecture Notes in Data Mining. University Of Tennessee, USA.
52. Ian, H.W. and Eibe, F. (2005).Data Mining Practical Machine Learning Tools and Techniques. 2nd ed., University of Waikato, New Zealand.
53. Liu, H. and Motoda, H. (1998). Feature Selection for knowledge Discovery and Data Mining Available URL: http://www.databaseheadquarters.com/bookstore/management2/079238198XAM US141630.shtml
54. Larose, D. T. (2005). Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.
55. International Development Research Centre. (2002). Population and Health in Developing Countries. INDEPTH Network 2002, Vol. 1, Ottawa, ON, Canada K1G 3H9. Available URL: http://www.idrc.ca.
56. George, A. (2004). Application of Data Mining in Medical Applications. MSc. Thesis, University of Waterloo, Ontario, Canada.
57. Oprean, C. (2011).Towards user assistance in Data Mining. MSc. Thesis, University of Waterloo, Ontario, Canada.
58. Kaur, H. and Krishan, S. W. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. Journal of Computer Science, New Delhi, India.
59. Jing-song, L. et. al. (2007). Data Mining in Hospital Information System. MSc. Thesis, Zhejiang University, China.
60. Uddin, J. et. al. (2008).Child Mortality in a Developing Country: A Statistical Analysis. Journal of Applied Quantitative Methods, Sylhet, Bangladesh.
61. Margaret, T. et. al. (2005). Oracle Data Mining Concepts. 10g Release 2 (10.2), CA, USA.
62. Nitesh, C. et. al. (2007).SMOTE Boost: Improving Prediction of the Minority Class in Boosting. Journal of Artificial Intelligence Research, University of Minnesota, USA.
63. Cios, K. and Kurgan, L. (2005). Trends in data mining and knowledge discovery. Springer Verlag, London, UK.

64. Chapman, P. et. al. (1999, 2000). CRISP-DM 1.0 Step-by-step data mining guide SPSS Inc., U.S.A.
65. Bao, H. (2003). Knowledge Discovery and Data Mining Techniques and Practice. http://www.netnam.vn/unescocourse/knowlegde/3-1.html.
66. Koliastasis, D. and Despotis, D. K. (2004). Rules for Comparing Predictive Data Mining Algorithms by Error Rate. OPSEARCH, VOL. 41, No. 3, 2004, University of Piraeus, Greece.

ANNEX III

A Partial J48 DT Generated for BRHP DSS Dataset

=== Run information ===

Test mode: split 90.0% train, remainder test

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

J48 pruned tree

AGE <= 2.997293

| TIMEX <= 37.708323

| | OU = NO

| | | IN = NO

| | | | TIMEX <= 4.944537: DEAD (1567.0/10.0)

| | | | TIMEX > 4.944537

| | | | | REL = CH

| | | | | | WINDOWS = UK: DEAD (18.0/1.0)

| | | | | | WINDOWS = YE

| | | | | | | TIMEX <= 21.990731: DEAD (156.0/10.0)

| | | | | | | TIMEX > 21.990731

| | | | | | | | TIMAD_1 <= 2.561074

| | | | | | | | | OXEN = UK

| | | | | | | | | | RELIG = OC: ALIVE (5.0)

| | | | | | | | | | RELIG = MU: ALIVE (11.0)

| | | | | | | | | | RELIG = UK: DEAD (2.0)

| | | | | | | | | | RELIG = CH: ALIVE (0.0)

| | | | | | | | | | RELIG = OT: ALIVE (0.0)

| | | | | | | | | OXEN = NO

| | | | | | | | | | TIMAD_1 <= 1.496079: ALIVE (7.0/2.0)

| | | | | | | | | | TIMAD_1 > 1.496079: DEAD (15.0/3.0)

| | | | | | | | | OXEN = SI: DEAD (10.0/1.0)

| | | | | | | | | OXEN = TW: DEAD (0.0)

| | | | | | | | TIMAD_1 > 2.561074: DEAD (25.0/1.0)

Number of Leaves: 215

Size of the tree: 327

Time taken to build model: 1.08 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	1904	97.491 %
Incorrectly Classified Instances	49	2.509 %
Kappa statistic	0.9486	
Mean absolute error	0.039	
Root mean squared error	0.1486	
Relative absolute error	7.9676 %	
Root relative squared error	30.0675 %	
Total Number of Instances	1953	

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.982	0.035	0.974	0.982	0.978	0.988	Alive
	0.965	0.018	0.976	0.965	0.97	0.988	Dead
Weighted Avg.	0.975	0.028	0.975	0.975	0.975	0.988	

=== Confusion Matrix ===

a	b	<-- classified as
1105	20	a = Alive
29	799	b = Dead