ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Predicting Under-Five Children Mortality Using Data Mining Techniques

By

Hailemariam Tafesse

To

Temtim Assefa (PHD)

October

2019

ACKNOWLEDGEMENTS

First of 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 glory, Amen!

My honorable gratitude extends to my advisors Dr. Temitm Assefa for the continuous support and guidance in the whole undertakings associated with my study. But special thanks go to my Sister W/rt. Workenesh Tafesse, without you this research would have not been a success. W/rt. Workenesh, you grow me by scarified your life, I never forget throughout my life with your helpful personality will always be a role in my heart.

I would also like to thank Central Statics Agency and Federal Ministry of Health for allowing me to carry out this research using the required data in addition to expert advice in the domain area.

My appreciation and thankfulness also go to my parents, brothers, and sisters, who morally and financially contributed to the completion of my study and A great deal of thanks should also be granted to my friends encouragement they rendered to me since the time of my admission to the postgraduate program.

Contents

List of Figures

List of Tables

Lists of Abbreviation

AI:	Artificial Intelligence
ARFF:	Attribute Relation File Format
AUC:	Area under the ROC
ANC:	Antenatal Care
BRHP:	Butajira rural health project (BRHP)
CRISP-DM:	Cross-Industry Standard Process for Data Mining
CSV:	Comma Separated Value
CSA:	Central Statistics Authority
DM:	Data Mining
DHS:	Demographic and Health Survey
EDHS:	Ethiopian Demographic and Health Survey
GUI:	Graphical User Interface
HSDP:	Health Service Development Program
ID3:	Interactive Dichotomizer 3
ICF:	International Confederation of Trade
KD:	Knowledge Discovery
KDD:	Knowledge Discovery in Databases
KDP:	Knowledge Discovery Process
KM:	Knowledge Mining
MDG:	Millennium Development Goal
NGO:	Non-Governmental Organization
ROC:	Receiver Operating Characteristics
SDGs:	Sustainable development Goals
SEMMA:	Sample, Explore, Modify, Model, and Assess
SMOTE:	Synthetic Minority Oversampling Technique
UNDP:	United Nation Development Program
UNICEF:	United Nations International Children's Fund
WEKA:	Waikato Environment for Knowledge Analysis
WHO:	World Health Organization

ABSTRACT

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.

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 EDHS-CSA 2016 datasets. This can greatly support for policy makers, planners, and healthcare providers working on the control of under-five children mortality in Ethiopia.

The methodology used for this research was a hybrid six-step Knowledge Discovery Process. The required data was received from DHS sites which covering the period 2016. The researcher used two popular data mining algorithms (Random Forest and PART Classifier) to develop the predictive model using a larger dataset (16,650 household). The researcher also used a 10-fold cross validation and 80% split test mode for data mining methods of the two predictive models for performance comparison purposes.

The results indicated that the Random Forest is the best predictor with 99.32% accuracy. Thus, the study results reveal that several socioeconomic, demographic and health related variables associated with under-five mortality. This analysis results indicate that the best attributes selected for under-five mortalities are place of residence, household wealth; number of antenatal (ANC) visits, level of education, employment, religion, experience on pregnancy termination, experience on pregnant drug taking, vaccination of child, and size of child are major predictors. Therefore, attention should give to these predictors to reducing the risk of child mortality.

The results from this study shows that applying data mining techniques could support in designing a predictive model for determining under-five mortality in Ethiopia. In the future, more classification studies by using a possible large amount of demographic and surveillance dataset records with epidemiological information and employing other classification algorithms, tools and techniques could yield better results.

Chapter One

Introduction

1.1. Background

Death is certain and it is a natural phenomenon that everyone has to face. The maximum age of living varies based on our living standards. Globally, life expectancy in 2015 was 70.8 years and an average women life expectancy is higher than men thus the overall female life expectancy is 73 years and male life expectancy is 69 years according to WHO and UN reports [1]. The gap between African life expectancy and European life expectancy has narrowed by 4.9 years since the year 2000 according to UN and WHO reports [1, 2]. In developed countries, the maximum living age expectancy was 80 and in developing countries maximum living age expectancy was 60 [2, 3].

According to World Health Organization (WHO) estimates, the decrease in deaths of child, maternal, infectious diseases and non-communicable disease would result in an increase in global average life expectancy of around 4 years by 2030 [1].

However, death will be the natural phenomena but child death should not be easily acceptable when we thought as human being. We bring children in this world with our consent so child never dies by preventable diseases [4]. The most horrible fact about child death occurred by diseases that are readily preventable or treatable with proven, cost-effective and quality-delivered interventions. Infectious diseases and neonatal complications are responsible for the vast majority of under-five deaths globally [4].

Under-five mortality rate (UFMR) reflects the socioeconomic, health, and environmental conditions in which a child lives and develops [5]. It defined as the probability of a child dying before attaining the exact age of 5years calculated per 1000 live births. Globally, UFMR has declined from 205 per 1000 live births in 1990 to 59 in 2015, which corresponds

to a 5% reduction [6]. However, the world could not achieve the Millennium Development Goal (MDG) 4 of reducing UFMR by two-thirds in 2015 than in 1990 [5]. In 2015, more than 80% of the total 5.9 million under-five deaths was estimated to have occurred mostly from preventable causes in developing countries [5].

Under-five mortality rate as a socioeconomic and health barometer has thus been included in the Sustainable Development Goals (SDGs) with a renewed target of reducing the overall global mortality rate to less than 25 per 1000 live births by 2030 [7].

Ethiopia has been able to make significant progress in improving the health of its population. The UFMR declined from 205 per 1000 live births in 1990 to 59 in 2015[5]. And further to 59 per 1000 live births in 2015[5], thus, achieving the MDG goal 4. Nonetheless, its current UFMR is still higher than that of some countries in the Africa region [5].

However, child mortality levels and trends in the developing countries become hazards. As a result of the absence of systematically organized registration of vital events in developing countries, adequate and reliable health information is often lacking [8, 9]. Therefore, Demographic Health Survey (DHS) studies can fill this gap of data inadequacy and problem of health care utilization in developing countries. This in turn to generate sound data on morbidity, mortality, and fertility through DHS [10, 11]. In Ethiopia, DHS were established and data collected on vital events like death, birth, and related research has been conducted for the last 32 years (1984-2016) [9]. This is mainly to track a limited and common set of key variables that deal about population dynamics and demographic trends. Therefore, DHSs has an approach to define key variables with their relationships and a developed system for collection, storage and analysis of surveillance data [12].

In sub-Sahara Africa, under-five mortality is one of the major concerns; and Ethiopian is one of those countries in the world, nearly half of children death under the age of five years occur during the first 28 days [2]. Therefore, the purpose of this research is to examine the determinant factors of under-five children mortality and build a model that predicts the under-five children mortality using data mining and designing the prototypes of an interface with Microsoft database.

Nowadays, the fast-growing, tremendous amount of data which is collected and stored in large and numerous data repositories has far exceeded human ability for grasp without powerful tools [13]. As a result, data collected in large data repositories infrequently visited. Consequently, important decisions often made not based on the information-rich data stored in data repositories rather on a decision maker's perception. This is due to decision makers lack the tools to extract the valuable knowledge embedded in the vast amounts of data [13].

In order to discover unforeseen relationship and propose solutions, DHSs and epidemiological tools are inefficient and unable to discover new and interesting patterns[14]. So data mining has evolved as a new technique and methods to evaluate, analyze, search and discover new patterns and relationships hidden in large database [14]. Therefore, it is important to explore child data from different perspectives. Moreover, it requires the application of different type of data mining techniques. Data mining is the process of extraction of previously unknown and potentially useful information and patterns from vast amount of collected data using different algorithms like classification and clustering [10, 14, 15]. Data mining is a repeated process to get valuable and substantial information from big data and it needs to integrate the efforts of human experts and machine. So to get the best result we need to balance this two [11, 16].

Data mining tasks are in general classified in to two main categories: predictive and descriptive data mining [17]. Predictive modeling is one of the data mining tasks that allow learning a mapping from an input set of vector measurements to a scalar output in which the training data consists of pairs of measurements. The goal of predictive modeling is to estimate (from the training data) a mapping or a function that can predict a value given an input vector of measured values and a set of estimated parameters for the model [18, 19]. The second category of data mining function is descriptive mining task used to characterize the general properties of the data in the database [20].

Thus, applying the data mining techniques intended to address diverse problem associated with child health and to extract useful knowledge from the DHS datasets. Exploring data mining technology to predict the risk of child mortality based up on Demographical Health Survey datasets gathered by CSA study considered as the main task to build predictive model after identifying socio-demographic and other relevant predictors that are associated with child mortality.

1.2. Statement of the Problem

In Ethiopia, Large efforts has been occurring to changing their names from developing countries to middle income countries however they have faced an enormous problem like economical and resource limitation. Ethiopia recognizes the central role of health care in improving the health of its citizens. So, the country is working on designing the Health Sector Development Plan (HSDP I,II,III,IV) to improve the health services and reduce child and Maternal mortality [18, 21]

Previously, Ethiopia is one of the highest child mortality countries rate among seven countries but now achieving MDGs before 2015 deadline with high decline in child mortality as stated in UN Inter-Agency group for child mortality estimation [1, 20].

However, most of the deaths caused by factors attributed to pregnancy and childbirth [16]. Ethiopia government builds health facility to provide health services for the community [20]. However, they don't have experience documenting and sharing data when the incident happened especially for child deaths [9, 14]. To fill this data gap, CSA carried out countrywide DHS in 2016 [8, 16].

This organization uses statistical tools to show summary figures of different variables involved in 2016 child mortality EDHS database [17]. However, the inefficiency of these tools needed the development of methods that are more powerful and techniques that used to study relationships and patterns through the large volumes of data collected.

Therefore, the fundamental research problem that demanded this research is that even if the country meets MDGs at 2015 still there is high under five-child mortality at a national level. Child mortality is high as compared to some other countries and still needs special emphasis to reverse the conditions. Among the mortality indicators, cause of death is one of the most highly problematic, especially for developing countries [22]. Tracking change in the basic outcome of child health is important for assessing progress, improving interventions, and driving further investment [23, 24].

Different studies has been conduct on child mortality using the Demographic Survey (DHS) data. Mahy [25] analyzed DHS data for the risks of death during four birth intervals, neonatal, infant, early childhood and under-five, and found that children from higher-order births (due to high fertility) are known to be at greater risk of dying during infancy and early childhood. Short inter-pregnancy intervals a characteristic of high-fertility counties also substantially contributes to child mortality, low birth weight, preterm birth and small size for gestational age. Abera [26] 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 observed

in female children than in males; moreover, multiple births were at increased risk of dying than singleton.

Another study conducted by Tanja et al. [27] reveals that potential markers of epidemiological transition like literacy, source of water, distance from town, house ownership, age group, sex, and period are risk factors for mortality. The global pattern of child morality also indicates that that socioeconomic condition is an important determinants of child survival i.e. the income levels of the populations has direct relationship with child survival [27, 28].

Consequently, there are studies using data mining to explore patterns from BRHP data. Shegew [29] applied data mining techniques to predict the risk of child mortality in the area. Amanuel [30] also conducted a research in the area by using data mining techniques in order to predict household health seeking patterns using BRHP dataset. His intention was to develop a model that identifies risk factors and patterns of household health seeking behavior at Butajira district. Recently, Taddesse [31] also applied data mining techniques to discover knowledge to gain insights in to vital statistics using 18 years BRHP data.

Though different studies have been conduct using epidemiological tools and data mining technology, all the available knowledge in the area are insufficient to solve the problem of the child mortality, as it is becoming an important indicator population health.

Thus, this research aimed to examine the determinant factors of under-five children mortality and to design a model that predicts the under-five children mortality using data mining method.

This research plan to address the following research questions:

→ What are the determinant factors that predicts under five-child mortality in Ethiopia?

→ Which predictive modeling algorithms are suitable for determining mortality status of under-five children?

→ How we build and evaluate under five-child mortality predictive model for designing knowledge sharing platform?

1.3. Objective of the Study

1.3.1. General Objective

The general objective of this study is to design a predictive model that determine under-five child mortality.

1.3.2. Specific Objectives

In order to achieve the general objective, the following specific objectives are stated.

- To undertake data preprocessing to generate error free data for model building.
- To select appropriate classification algorithm to build a model that predicts child mortality
- To construct child mortality predictive model using the selected classification algorithms.
- To design the Knowledge platform and evaluate using user acceptance testing.

1.4. Significance of the Study

Data mining used to make informed decision by extracting hidden knowledge and pattern from large volume of data.

Therefore, the significance of the study is that it provides new knowledge on factors that predicting under-five child mortality. This will help health administrative practitioners to

improve the quality of services. In addition, the result of this research used as an input in national health care policy revision.

Moreover, the result of the study used as a benchmark for interested researchers to explore the issues in the area of under-five mortality. The model used by policy maker, planner and FMOH to take any necessary decision to improve child health care services.

1.5. Scope and Limitation of the Study

This study used secondary data collected by central statistical authority (CSA), i.e. the Ethiopian demographic and health survey of the year 2016. The research aimed to apply data mining techniques for discovering significant knowledge using DHS data and build a model that predicts under five children mortality through identifying dominant factors related to cause of under five children death in Ethiopian.

The research also restricted to use the socio-economic and demographic factors to develop the model. In addition, based on the model the user interface designed for the deployment of the results to reveal the knowledge. However, the major limitation of this research is, due to time limitation the researcher is unable to apply association rule discovery techniques to investigate the internal association exists among the different variables considered in this research. Furthermore, the study did not incorporate the regional socio-economic and demographical factors rather it focused at national level.

1.6. Ethical Consideration

For this study, the personal identification (name and/or ID) does not require, beside, the study anticipated fully for academic purposes for partial fulfillment of M.Sc degree of information Science. Therefore, for this study the researcher will use EDHS 2016 data, which is publicly available or made available when request from the following links

(https://www.dhsprogram.com/). This data collected by CSA authorization level so no data collected in individual level rather the researcher used secondary CSA data. This research outcome anticipated to contribute to the health promotion and policy changes in urban and rural Ethiopia level.

1.7. Organization of the study

This research report 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, scope the study, and Ethical clearance for the study. Chapter 2 and 3, review the DM technology and the methods for mining under-fives mortality on EDHS 2016 data respectively. The concepts pertaining to the DM technology and its application in the problem reviewed in chapter two. Moreover, the application of DM in healthcare and related works on DM reviewed in this chapter two. Chapter 3 is dedicated for the discussion of basic issues, tools, techniques and algorithms that can be relied on the under-fives mortality at EDHS 2016 (the problem area). Chapter 4 explains the Business Understanding and Data Preprocessing used in this research. Chapter 5 presents the experimentation and analysis phase of the study at hand. As well as the results of the classification experiments discussed here. Chapter Six present the demonstration of prototype of child mortality using visual studio and MS-Access database. Finally, chapter seven provides conclusion of the research, and presents recommendation for future work.

Chapter Two

Literature Review

2.1. Overview

This section deals with literatures conducted around child mortality and data mining techniques that involves on child mortality found in the literature. Generally, the literatures are present here to show that the selected approaches have evidence. So following on this will define mainly the child mortality in global perspective and country (Ethiopian) perspective and then describe the selected approaches, the two data mining techniques predictive and descriptive were used as method.

2.2. Under-five child Mortality

According to EDHS 2016 [16], under-5 mortality is the probability of dying between birth and the fifth birthday; which means it divide in to two big categories infant mortality and child mortality.

Infant mortality is the probability of dying before celebrating the one year birthday; it includes the neonatal mortality (the probability of dying within the first month of life) and post-neonatal mortality [16, 19] (the probability of dying after 30days of life but before the first birthday). Infant mortality rate (IMR) defined as the rate for a given region is the number of children dying under one year of age, divided by the number of live births during the year, multiplied by 1,000 [16, 19].

Child mortality is the probability of dying between the 12 months and the fifth birthday. Child Mortality rate (CMR) defined as the rate for a given child mortality is the number

of children dying between one year of age and five years, divided by the number of live births during the year, multiplied by 1,000.

A child's death is very hard for the parents and their relatives; so that many child deaths not reported, and it is difficult to track to bring rapid changes. Therefore, Reduction of child mortality is the fourth priority in the united nation millennium development Goals (MDGs). However the MDGs replaced in 2015 by the sustainable development goals (SDGs) [32]. According to the estimate developed by the WHO, in 2015, there were 5.9 million children died before their fifth birthday, 45% (2.7 million) of the babies die every year in their first month of life that means child died during the first 28 days of life [4, 33, 34]. Substantial global progress has been making in reducing child deaths since 1990.

The number of under-5 deaths worldwide has declined from 12.7 million in 1990 to 5.9 million in 2015 – 16,000 every day compared with 35,000 in 1990. Since 1990, the global under-5 mortality rate has dropped 53%, from 91 deaths per 1,000 live births in 1990 to 43% in 2015 [33].

Between 1990 and 2015, 62 of the 195 countries with available estimates met the MDG4 target of a two-thirds reduction in the under-5 mortality rate. Among them, 24 are low- and lower-middle income countries. Despite these gains, progress was insufficient to reach SDGs globally and in many regions. Currently, 79 countries have an under-5 mortality rate above 25 deaths per 1000 live births, and 47 of them will not meet the proposed SDG target of 25 deaths per 1000 live births by 2030 if they continue their current trends in reducing under-5 mortalities. Among these 47 countries, 34 are in sub-Saharan Africa. The acceleration needed to reach the goals in those 47 countries is substantial – 30 countries must at least double their current rate of reduction, and 11 of those 30 countries must at least triple their current rate of reduction. Wide gaps in child mortality across sub-groups or areas within countries have been document, warranting a

call for an equity-focused approach to reducing child mortality[2, 35]. Children are at greater risk of dying before age 5 if they are born in rural areas, poor households, or to a mother denied basic education[4, 32].

2.2.1. Causes of Child Mortality

Child survival is a field of public health concerned with reducing child mortality. Child survival interventions designed to address the most common causes of child deaths that occur, which include diarrhea, pneumonia, malaria and neonatal conditions. So far, there is six conditions account for about 70% of all child deaths: acute lower respiratory infections, mostly pneumonia (19%), diarrhea (18%), malaria (8%), measles, (4%), HIV/AIDS (3%), and neonatal conditions, mainly pre-term birth, birth asphyxia, and infections (37%). However, more than half of under-5 child deaths are due to diseases that are preventable and treatable through simple, affordable interventions[36]. Strengthening health systems to provide such interventions to all children will save many young lives [32, 37]. Malnourished children, particularly those with severe acute malnutrition, have a higher risk of death from common childhood illness such as diarrhea, pneumonia, and malaria. Nutrition-related factors contribute to about 45% of deaths in children under-5 years of age [4, 32].

Ending preventable child deaths can be achieve by providing immediate and exclusive breastfeeding, improving access to skilled health professionals for antenatal, birth, and postnatal care, improving access to nutrition and micronutrients, promoting knowledge of danger signs among family members, improving access to water, sanitation, and hygiene and providing immunizations. Many of these lifesaving interventions are beyond the reach of the world's poorest communities.

Why are child death rings still high?[4]

- Poor families are often unable to obtain even the most basic health care for their children. Poor or delayed care seeking contributes to up to 70% of all under-five child deaths.
- Of the 12 countries where more than 20% of children die before their fifth birthday, nine have suffered a major armed conflict in recently.
- Countries with weak and fragile health systems have not been able to provide effective child survival strategies that are crucial to reduce under-five child deaths, and especially neonatal deaths. Basic health services have been lacking as well as nutrition, water supplies and sanitation facilities.
- Almost half a million deaths each year due to malaria in children under-five in sub-Saharan Africa could have been prevented with the use of insecticide-treated bed nets, shown to reduce under-five mortality rates by up to 20%.

2.2.2. Child Health in Ethiopia

In the last twenty-five years, the world show significant changes in improving child survival, the main progress shown in reducing child mortality throughout the world and this change saved millions of lives of children under-5 age. The incremental speed in 2000-2015 periods compared with the 1990-2000s period hopeful to the future. Therefore, the annual rate reduction in the under-5 mortality rate has shown 1.8 percent increment in 1990-2000 to 3.9 percent in 2000-2015 [4, 32, 37].

Ethiopian is one of the sub-Saharan countries working to shift from lowest income country to middle-income country. However, previously, child health status in Ethiopia is one of the worst in the worlds; so the country is considered as high maternal and child mortality. According to UNICEF report, "Ethiopian child is 30 times more likely to die by his or her fifth birthday than a child in Western Europe" [34]. However, this figure seems like changed, relative to the progress shown throughout the world in the past 25 years as UNICEF report 2015. Ethiopia is one of the countries from 62 countries, which have reduced their under-5 mortalities by two-third or more and achieved the MDGs4 target

set in 2000. The child mortality rate was estimated at 20/1,000 children surviving to age 12 months according to the 2016 EDHS and infant mortality rate was 48/1000 [16]. While, the overall under-5 mortality rate was 67 deaths per 1,000 live births. The neonatal mortality rate was 29 deaths per 1,000 live births, and the post-neonatal mortality rate was 19 deaths per 1,000 live births. The 2016 EDHS findings further indicate that all childhood mortality rates have declined over time. For example, the under-5 mortality rate has declined from 166 deaths per 1,000 live births 2000 EDHS to 67 deaths per 1,000 live births 2016 EDHS [16].

2.3. Overview of Data Mining

As we observed that in every field of working area there is an increasing of the accumulation of raw data day to days and this is creating an opportunity and challenges for the process of knowledge discovering. The challenge associated with the largeness of the data size related to the limited processing capabilities of prevailing statistical tools, which lead to a demand for better methods to deal with the large volume of data. However, challenges are not the only thing that large data bases have come up with, opportunities are also associated with them i.e. they possess patterns and hidden information that represent interesting and hidden knowledge. Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database. The discipline concerned with this task has become known as data mining [38, 39].

Data mining defined by different scholars and it has several definitions. The one which is most frequently used by scientific community is that: Data mining is an activity that extracts new nontrivial information contained in large databases in order to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine learning, statistics and database technologies [40].

Data mining is applied by many fields like agriculture, education, business etc. [13]. It has also multi-effects in medicine; for example, it can help healthcare industry to make customer relationship management decisions, physicians to identify effective treatments and best practices, and patients to receive better and more affordable healthcare services [41].

In healthcare activities now numerous and complex amounts of data are generated and stored. Thus, the application of traditional tools become inefficient to discover useful information from such a vast data. This, in turn calls data mining that enables to explore the information buried in the data and creates models to find hidden patterns in large and complex collections of data which overcomes traditional methods of data analysis because of the large number of attributes and the complexity of patterns[42].

2.4. Methodology of Data Mining Research

Process means the task done to achieve something to fulfill our desires. In addition, related with this, there is confusion between process and methodology.

Process is a series of actions or operations conducing to an end. This means that, it represented by a sequence of steps executed in order to produce a certain result and it is especially useful in any sort of repetitive task, making the same thing over and over and over again certainly lends itself to a particular set of steps that ought to be optimized.

Methodology a body of methods, rules, and postulates employed by a discipline a particular procedure or set of procedures so we can define, 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 [43].

DM is one of the most important steps in the KDD process. Therefore, it can be considered that the heart of the KDD process. KDD is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [13, 44].

Nowadays, with the explosion of information, DM has become one of the top ten emerging technologies that will change the world[45]. So, if we are used one of the data mining procedures and planned carefully, then we can get successful result at the end[38].

2.4.1. Knowledge Discovery in Database (KDD)

Traditional data analysis techniques facilitate useful data interpretations and can help to generate important insights into the processes behind the data. These interpretations and insights are the ultimate knowledge sought by those who build databases. Yet, such knowledge created by these tools, but instead has to be derive by human data analysis.
In efforts to satisfy the growing need for new data analysis tools that will overcome the above limitations, researchers have turned to ideas and methods developed in machine learning. The field of machine learning is a natural source of ideas for this purpose, because the essence of research in this field is to develop computational models for acquiring knowledge from facts and background knowledge. These and related efforts have led to the emergence of a new research area, frequently called data mining and knowledge discovery.

There is confusion about the exact meaning of the terms "data mining" and "KDD." KDD proposed in 1995 to describe the whole process of extraction of knowledge from data. In this context, knowledge means relationships and patterns between data elements. "Data mining" used exclusively for the discovery stage of the KDD process.
The first KDD process was proposed by Fayyad in 1996 [44]. This process consists of several steps that can be execute iteratively. KDD has been more formally defined as it is non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is the process of knowledge discovery while data

mining is a technique applied for knowledge discovery considered as just a step in the entire process [44]

As shown in Figure 2.1, the KDD process consists of five steps: Data Selection, Data Pre-processing, Data Transformation, Data Mining and Interpretation/Evaluation [46, 47].

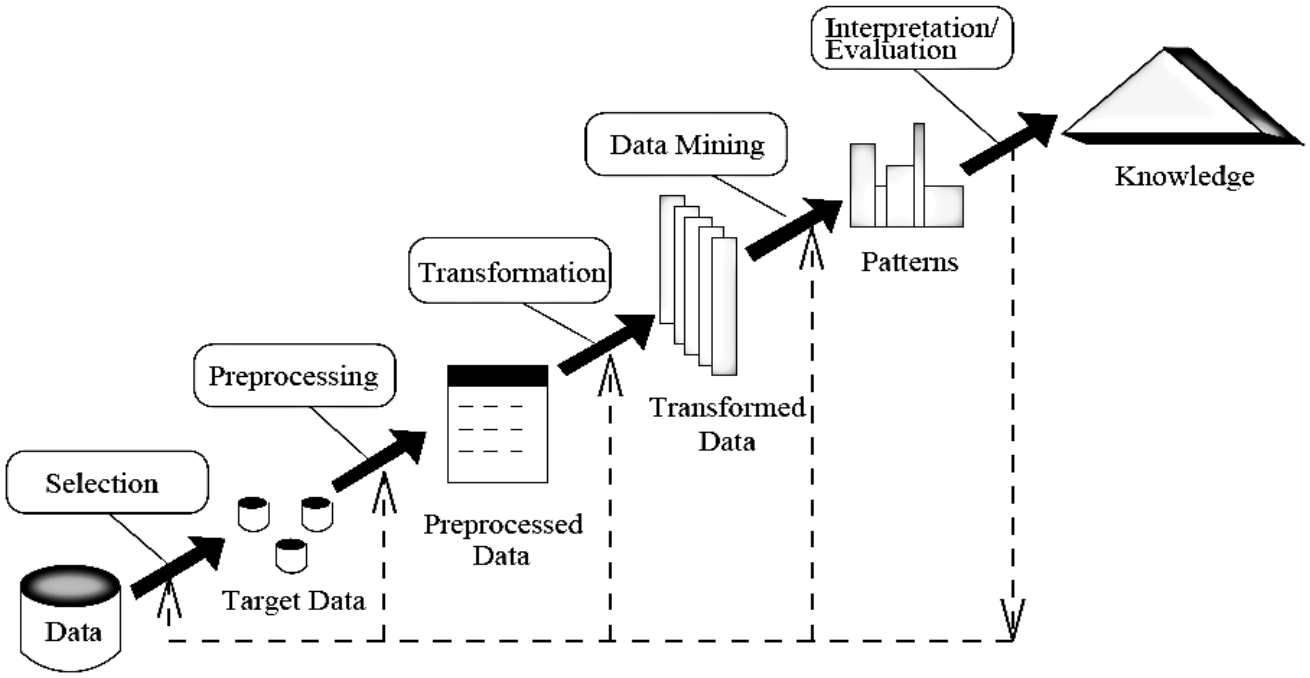


Figure 2. 1: KDD Process Model, adapted from Fayyad et al. (1996) [48]:

The KDD process consists of the following five steps:

1. **Select a target data set:** The initial step based on data needed for the DM process may obtained from many different and heterogeneous data sources.
2. **Data preprocessing:** In this step, the data to be use 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 added and/or eliminated from the target data. Data from different sources must converted into a common format for processing.

4. **Data mining:** A best model for representing the data 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 discovered is both useful and interesting.
6. Another important step not contained in the KDD process is goal identification. The focus of this step is on understanding the domain being consider for KD.

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

The CRISP-DM Process Model is the leading method used by data miners; this result reflected that in 2002, 2004, 2007 and 2014 at the time polls conduct [49]. The only other data mining standard named in these polls was SEMMA[50]. However, 3–4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008 [50, 51].

CRISP-DM is a non-proprietary, documented and freely available data mining process model. It developed by the industry leaders and the collaboration of experienced data mining users, data mining software tool providers and data mining service providers. CRISP-DM is an industry-, tool-, and application-neutral model created in 1996 [51-53]. Special Interest Group (CRISP-DM SIG) formed in order to further develop and refine CRISP-DM process model to service the data mining community well. CRISP-DM version 1.0 was presented in2000 and it is being accepted by business users [51-53].

CRISP–DM is one of the most widely used methodologies in extraction of knowledge which has a life cycle consisting of six phases which is an iterative and adaptive process[11], as depicted in Figure 2.2.

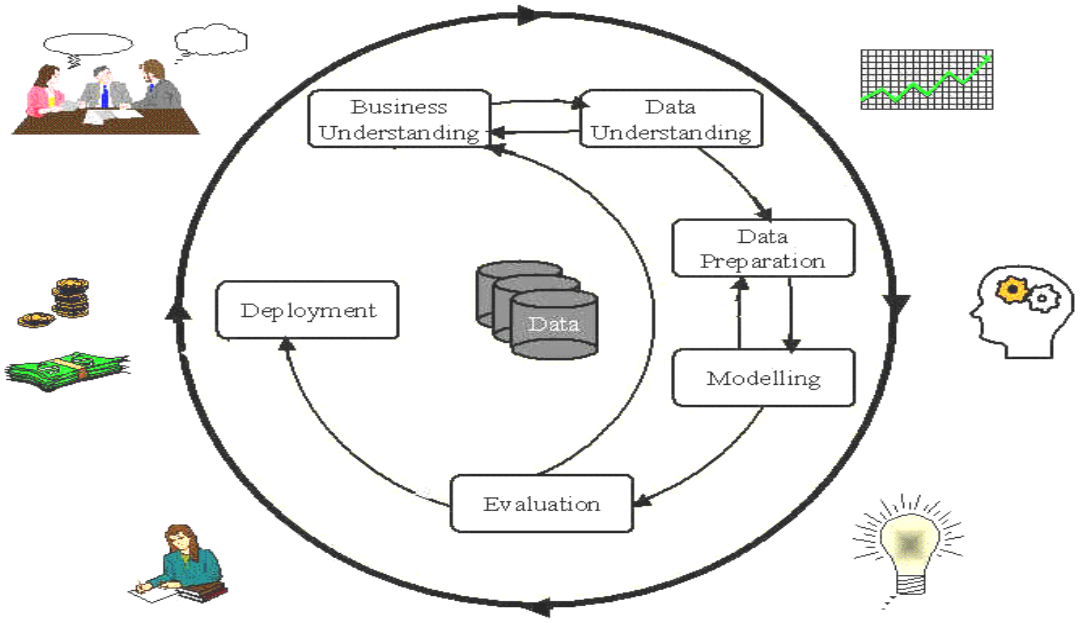


Figure 2. 2: The CRISP-DM knowledge discovery Process Model Adopted from Chapman et al. (2000) [52]

Phase-1:

Business understanding-this initial 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;

Phase-2:

Data understanding-the data understanding 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;

Phase-3:

Data preparation-the data preparation phase covers all activities to construct the final dataset from the initial raw data;

Phase-4:

Modeling-in this phase, various modeling techniques selected and applied and their parameters calibrated to optimal values.

Phase-5:

Evaluation at this stage the model (or models) obtained more systematically evaluated and the steps executed to construct the model reviewed to be certain it properly achieves the business objectives.

Phase-6:

Deployment creation of the model 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 [50].

2.4.3. SEMMA

In order to apply successfully, the data mining solution must viewed as a process rather than a set of tools or techniques. In addition to the CRISP-DM there is yet another well-known methodology developed by the SAS Institute, called SEMMA. The acronym SEMMA stands for sample, explores, modify, model, assess. Beginning with a statistically representative sample of your data, SEMMA intends to make it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and finally confirm a model's accuracy [50].

A pictorial representation of SEMMA is given in Figure 2.3 [50].

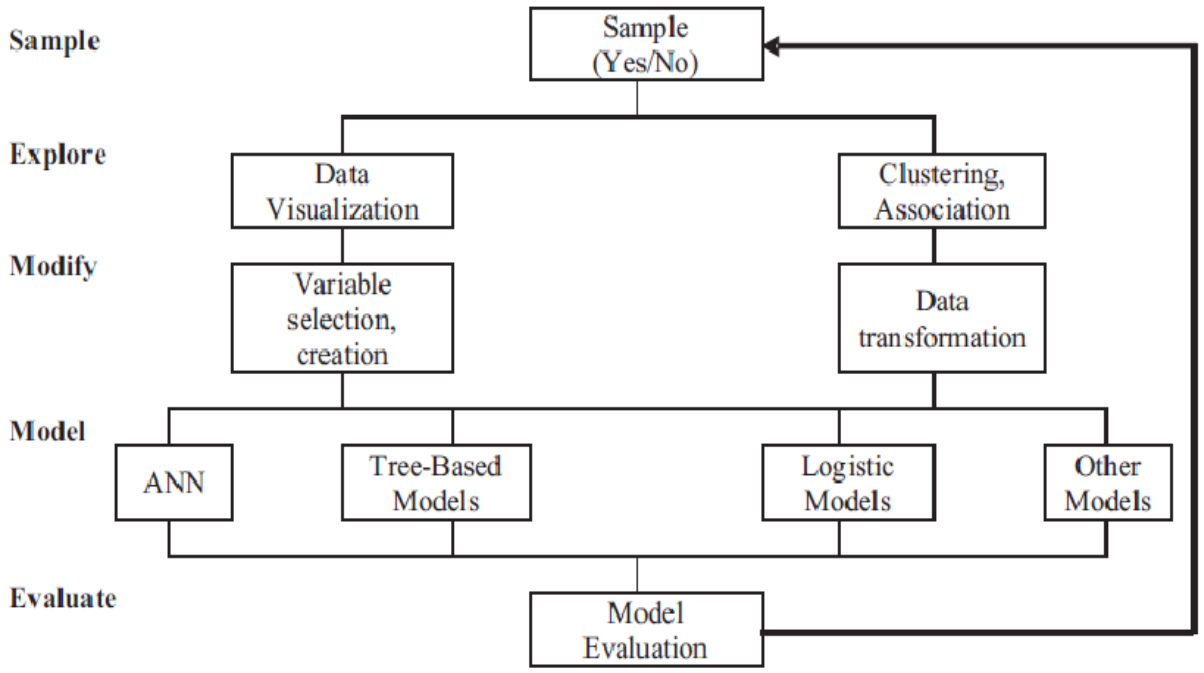


Figure 2. 3: Schematic of SEMMA (original from SAS Institute)[50]

Data Mining Process Steps in SEMMA Process are listed and shortly described below [50, 54].

Step 1 Sample:

This is where a portion of a large data set (big enough to contain the significant information yet small enough to manipulate quickly) extracted. For optimal cost and computational performance, some (including the SAS Institute) advocates a sampling strategy, which applies a reliable, statistically representative sample of the full detail data. In the case of very large datasets, mining a representative sample instead of the whole volume may drastically reduce the processing time required to get crucial business information. If general patterns appear in the data as a whole, these will be traceable in a representative sample.

Step 2 Explore:

This is where the user searched for unanticipated trends and anomalies in order to gain a better understanding of the data set. After sampling your data, the next step is to explore them visually or numerically for inherent trends or groupings. Exploration helps refine and redirect the discovery process. If visual exploration does not reveal clear trends, one can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering.

Step 3 Modify:

Based on the discoveries in the exploration phase, one may need to manipulate data to include information such as the grouping of customers and significant subgroups, or to introduce new variables. It may also be necessary to look for outliers and reduce the number of variables, to narrow them down to the most significant ones. One may also need to modify data when the "mined" SEMMA data change. Because data mining is a dynamic, iterative process, you can update data mining methods or models when new information is available.

Step 4 Model:

This step4 is the user searches for a variable combination that reliably predicts a desired outcome. Once you prepare your data, you are ready to construct models that explain patterns in the data. Modeling techniques in data mining include artificial neural networks, decision trees, rough set analysis, support vector machines, logistic models, and other statistical models such as time series analysis, memory-based reasoning, and principal component analysis. Each type of model has particular strengths, and is appropriate within specific data mining situations depending on the data.

Step 5 Assess:

This is where the user evaluates the usefulness and the reliability of findings from the data mining process. In this final step of the Datamining process user assesses the models to estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set put aside (and not used during the model building) during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, you can test the model against known data.

2.4.4. Hybrid Model

One of the most important applications in DM is hybrid modeling. A hybrid model is a collection of models of different types[55]. There are different reason hybrid data models built, some of them are because of the depth of understanding of the modeled process, accessibility of the data and specific drawback of application. The need for hybrid has been motivate by numerous engineering and medical applications. A typical process, e.g., a semiconductor manufacturing process or disease management process, involves stages that well understood due to available models and stages that are only loosely know. This lack of process knowledge is likely behind unwanted events in industrial processes (e.g., products below expected quality level) and in medicine (e.g., premature patient's death).

A hybrid approaches may be considered in determining a suitable goal for DM. Process model is the set of tasks to be performed to develop a particular element (outputs and inputs). The major motives for announcing process models is to formalize knowledge discovery projects within a common framework, a goal that will result in cost and time savings, and will improve understanding, success rates, and acceptance of such projects. The models emphasize independence from specific applications, tools, and vendors. KDD

process has a process model component because it established all the steps to be taken to develop a DM project[47].

The KDP model consists of a set of processing steps to be follow by practitioners when executing a knowledge discovery project. The model describes procedures that are perform in each of its steps. It is primarily use to plan, work through, and reduce the cost of any given project.

Since the 1990s, several different KDPs have been develop. The initial efforts led by academic research but quickly followed by industry. The first basic structure of the model was proposed by Fayyad et al [44]. In addition, later improved/modified by others. In 1996, the foundation for the process model was laid in a book entitled Advances in Knowledge Discovery and Data Mining [44]. All the proposed models also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process [47].

So far there is Five KDP models, suggested together-with the nine-step model by Fayyad et al [44], the eight-step model by Anand and Buchner, the six-step model by Cios et al., the five-step model by Cabena et al., and the CRISP-DM model were introduced [46, 47]. Each model has its strong and weak points, based on its application domain and particular business objectives. One thing we need to give attention is the data preparation steps, which are time taking portion of the KDP.

Knowledge discovery process Model categorized in to three generations:

1. **First generation systems: -**provided only one data mining technique, such as a decision tree algorithm or a clustering algorithm, with very weak support for the overall process framework. They were intending for expert users who already had an understanding of data mining techniques, the underlying data, and the knowledge sought. Little attention given to providing support for the data analyst, and thus the first knowledge discovery systems had very limited commercial success. The general research trend focused on the development of new and

improved data mining algorithms rather than on research to support other knowledge discovery activities.

2. **Second-generation systems: -**called suites, were develop in the mid-1990s. They provided multiple types of integrated data analysis methods, as well as support for data cleaning, preprocessing, and visualization.
3. **Third-generation systems: -** developed in the late 1990s and introduced a vertical approach. These systems addressed specific business problems, such as fraud detection, and provided an interface designed to hide the internal complexity of data mining methods. Some of the suites also introduced knowledge discovery process, to guide the user's work.

2.5. Data Mining Techniques

There is an increment of the volume of data and the size of the computer, so DM techniques provide a power to handle those data. However, the objective of DM varies based on the projected systems. Different methods and techniques needed to find different kinds of patterns. Data mining techniques used in many research areas, including health, telecommunication, marketing, financing, cybernetics manufacturing, etc. DM can help with data reduction, exploration, and hypothesis formulation to find new patterns and information in data that exceed human information processing limitations[56, 57].

In general, the primary data mining tasks categorized in to two[48]: Descriptive and predictive. Predictive data mining tasks focused on the value of one variable or fields to be predicted from the unknown or future values of other variables[48]. Descriptive mining tasks involves on presenting patterns that describe the underlying data. (I.e. description focuses on finding human-interpretable patterns describing the data)[48, 58]. As definition this two categories defined in separately but the boundaries of between the two were not clearly specify. (I.e. some of the predictive models can be descriptive, to the degree that they are understandable, and vice versa), the distinction is useful for understanding the overall discovery goal. The relative importance of prediction and description for particular data mining applications can vary considerably [48, 59, 60].

2.5.1. Predictive Data Mining Techniques

Predictive analytics uses data-mining techniques in order to make predictions about future events, and make recommendations based on these predictions. Moreover, predictive modeling is a technique that involves the use of 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 [61]. Predictive models were build, or trained, using data for which the value of the response variable already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results [62].

There are a number of predictive techniques, among which classification is one of them. The aim of classification techniques is to classify the determinants that indicate the group to which each case belongs [48, 63, 64]. This pattern can be used both to understand the existing data and to predict how new instances will behave [65]. Moreover, it is a technique used to predict group membership for data instances by assigning previously unseen records a class as accurately as possible. It is said to be the process of finding a model or function that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown [13]. The derived model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as tree (decision, random forest trees), decision lists (IF-THEN, PART rules), mathematical formula, semantic network etc. [13]. Each technique serves as learning algorithm by identifying a model that best fits the relationship between the attributes set and the class level of the input data. After having an accepted accuracy level, one can use the model for classification of new instances.

.

2.5.1.1. Classification by Tree Based Classifiers

Classification is the process, which assigns a specific item to one of the categories or classes specified, based on its features or properties. In machine learning, classification considered as a task to predict the value of one or more outcomes. The real task in classification is to find a relationship between features and its associated classes. There are various categories of classification which include linear classifiers, support vector machines, quadratic classifiers, kernel estimation, decision trees and neural networks [66]. Among these, tree based classifiers are commonly used for developing prediction algorithms for a target variable. Such classifiers construct a root node with a population of branches which comprise of internal nodes and leaf nodes [67].

Tree classifiers aim to partition dataset into groups of similar nature. They said to be very effective methods of supervised learning, which lead to generate unique solutions. In cases where impurity exists in the data and where there are traces of one class oversteps into another, tree classifiers are best suited. Unlike linear models, they map non-linear relationships quite well. It has many advantages like tree based is Comprehensive behavior, No need for tuning of parameter set, Easy to interpret, and Easily deal with outliers [67].

Tree algorithms extensively used for classification of various tasks in diverse domains. It include characterizing human behaviors, teaching, production of agricultural products, astronomical object study, financial, musical and medical study and so on, there are various applications of tree classifiers which consider classification of various instruments and speech/music classification and segmentation [68].

Random Forest is a classifier, which comprises the decision trees. It has shown that random forest performs equally well or better than other methods on a diverse set of problems. It has been widely used in classification problems as diverse as bioinformatics, medicine transportation safety and customer behavior [67].

Random forest offers a useful feature that improves our understanding of a classification problem under scrutiny. It gives an estimate of the importance of each attribute for final prediction. It is often use for analysis when both classifier and identification of important variables are goals of the study. Random forest collects votes from different decision trees, which are randomly select from training set data, and decides the final class of test data. This is helpful for finding accurate results because a single tree might lead to a noise, but a set of decision trees will reduce the noise [67].

2.5.1.2. Rule based classification

Though the decision tree is a widely used technique for classification purposes, another popular alternative to decision trees is classification rules which can be expressed as paths IF-THEN rules so that humans can understand them easily [69]. A rule-based classifier uses a set of IFTHEN rules for classification; it is a relationship between antecedent, and consequent i.e. an expression of the form IF condition THEN the conclusion. The algorithm decision tree is the best known method for deriving rules from classification trees [61]. For example, one could have the following set of rules to classify the weather condition. If temperature < 50°F, then weather = cold. If temperature > 50°F AND temperature < 80°F, then weather = warm. If temperature > 80°F, then weather = hot [70]. Although any of the logical expressions are allowed, preconditions are usually connected with the AND operation. The advantage of IF-THEN rule is the rules are order independent i.e. regardless of the order of rules executed, the same classification of the classes is possible to reach [70]. The challenges is the generated rules are often more complex than necessary and contain redundant information and the rules generated this way may be unnecessarily complex and incomprehensible [70].

2.5.2. Descriptive Modeling Techniques

Descriptive model is used to present the summary of data in descriptive ways (numerical or graphical) to be understood and interpreted by human [71]. The knowledge of the people about the products or process to describe what happening in the complicated database is important [72]. The following are the major techniques used by this method:

2.5.2.1. Clustering

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures.[13, 29, 65] According to them cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering[29]. In this context, different clustering methods may generate different clustering is on the same dataset. The partitioning is not perform by humans, but by the clustering algorithm. Hence, clustering is useful in that it can lead to the discovery of previously unknown groups within the data.

Clustering has its roots from many areas, including DM, statistics, biology, and machine learning[13] and applied in many applications areas, including pattern recognition, image processing, data analysis and marketing research[55, 65].

K-means, K-Medoids and others clustering analysis tools are built in several statistical packages (like SPSS, S-plus and SAS)[55]. Clustering can also be used for outlier detection[13] some of the techniques used for clustering are categorized as follows: 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[13]. Now the time clustering analysis is highly applied topics in data mining research, due to increasing the amounts of data

collected in databases. And the Clustering analysis focus only on the attribute associated with a given clusters[13, 55, 73].

2.5.2.2. Association Rule Discovery

Association rule is one of the major techniques of data mining which is identify rules about items that appear together in an event such as a purchase transaction, relational databases and other repositories [74]. This technique can help us extracting some unknown patterns in the datasets that can used to detect variables within the data and the concurrence of different variables that appear very frequently in the dataset. This rules are used in Market-basket analysis, product clustering and catalog designing are some of examples for Association discovery [13, 55, 73].

An association rule describes an interesting relationship among different attributes. A Boolean association involves binary attributes, a generalized association involves attributes that are hierarchically related, and a quantitative association involves attributes that can take on quantitative or categorical values [13, 74]. Association rule mining algorithms, mining associations from various types of data (e.g. numerical association rules, temporal association rules), filtering techniques for association rule mining, and interestingness measures for association rules are the main concepts in association rule mining. Besides, there are several extended association rules that are different traditional association rules. These extended association rules include peculiarity rules, informative rules, optimized association rules, representative association rules, high-utility rules, positive and negative association rules, substitution rules, and linguistic association rules [59].

Interestingness of patterns needs to be measured to make sense that patterns are easily understood by humans, valid, potentially useful, and novel [13]. In the case of classification

rules, we are generally interested in the quality of a rule set as a whole. It is all the rules working in combination that determine the effectiveness of a classifier, not any individual rule but, in the case of association rule mining the emphasis is on the quality of each individual rule [13].

2.6. Application of Data Mining in Healthcare

Healthcare systems are generating large and complex amount of data from their routine activity of service [75]. The data is too difficult to process and analysis using traditional statistics [15]. As Fayyad et al., defined, "traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive and highly subjective"[48]. Data mining, which is discovering knowledge form large datasets, helps the healthcare Sectors to analyzed and discover patterns in improving the quality of service delivery at healthcare.

Nowadays medical database is increasing in size in rapid ways. The database contains a valuable knowledge hidden on it; therefore, analyzing such data very important for healthcare decision makers and management. It has been widely accepted that the analysis is medical data can lead the development of healthcare by improving the performance of patient management tasks so exploring such using data mining approach has great value.

The analysis of healthcare data has two benefits [57, 76].

1. Support of specific knowledge-based problem solving activities through the analysis of patient raw data collected in monitoring.
2. Discovery of new knowledge that can be extract through the analysis of representative collections of example cases, described by symbolic or numeric descriptors.

Generally, when we see the healthcare data is huge however it found in unorganized ways because of they are stored in different repository systems i.e. excessiveness of the

data has no meaning unless the dissimilarity of the form and format handled properly [75]. This makes data mining of healthcare data different from others. Data mining in the healthcare data is an important input for healthcare management. Thus, to put the healthcare sectors in better position the healthcare managers can uses the identified knowledge for Decision making [75, 77]. For example, using the large item sets and decision tree classification method in data mining 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 [77]. Using the cluster technique in data mining 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 [77].

In addition to this, the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives, can also use the discovered knowledge. And also, data mining technology applied in identification of interesting patterns in infection control data so the healthcare managements concluded by support of experts "enhancing infection control with the data mining system is more sensitive than traditional inflation control surveillance, and significantly more specific" [57, 76].

Model building in data mining that helps to identify the trend of patients who are in high/low risk condition so this gives information to nurses care coordinators to prepare a head based on the risk level of the patients to improve the patients status and in the future to plan and prevents health problems [76].

The Following are some of the important areas of interests where DM techniques can be of tremendous use in health care management [78].

Data mining is used to support healthcare management [78]:

- ⇨ 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 healthcare
- ⇨ Health Insurance

Finally, data mining supports researchers to discover effective results and best practices in the field of healthcare sectors.

In healthcare sectors, data mining supports physician and decision makers to improve the quality of services. At all, the potential of data mining application in healthcare can be grouped as the Evaluation of treatment effectiveness; Management of Healthcare; Customer relationship management; and Detection of fraud and abuse etc.[79].

2.7. Related Works

In Ethiopia, only few studies have been undertake focusing on factors that affects children health service in the country and specially using data mining technology. One of the researches was conducted by Shegaw [29]. He investigated the potential applicability of data mining technology to predict the risk of child mortality based up on community-based epidemiological datasets gathered by the Butajira rural health project (BRHP) epidemiological study. The methodology was he used Neural network and decision tree data mining techniques were employed to build and test the models, by testing around 1100 sample datasets records. The best performing neural network model and decision tree classifier 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 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 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 nine input variables. This model had an accuracy rate of 93% (classified 102 of the 110 test cases correct) at a testing tolerance of 0.4 and tested with accuracy of 88 % (classified 97 of the 110 test cases correct) at testing tolerances of 0.2 and 0.1. As a result, the model, which trained by using the default parameters and the nine input variables predicted from out of ten new cases, the class of the 9 new cases correctly classified, and it misclassified one alive case as Died.

From decision tree classifiers, the best classifier achieved when the ruleset and adaptive boosting options used. 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 prediction performance of the decision classifier was evaluate by using the ten unseen cases that used to evaluate neural network models. As a result, the classifier predicted the outcome of the nine cases correctly, where as it wrongly predicted one alive case as died.

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 since it provides additional features such as simple and easily understandable rules that can be use by non-technical health care professionals as well as health care planners and policy makers.

Then Helen (2003)[80] attempted to investigate the application of data mining techniques to identify significant pattern for child labor using 2001 CSA data. Helen were used Aprior association rule algorithm to generate meaningful association rules from clustered and non-clustered selected datasets. Therefore, for this clustering purpose around 2398 records with 63 attributes were use from the datasets. The clustering algorithm, expectation maximization, applied using the 63 attributes used in experiment 2. Four clustering models were built by varying the number of clusters from 2 up to 5. The

cluster model, which according to the domain experts made good sense about child labor, segmented the records into five clusters. Among these five clusters, the third cluster, which contains 42.5% of the selected dataset, was chosen and given to the apriori algorithm of Weka. The association rule algorithm, apriori, generated its 10 best association rules with minimum coverage of 95% and minimum accuracy of 90%.

Dawit [81] and Zenebe [39] also attempted to develop a predictive model for maternal health care seeking and construct a model that predicts the under nutrition status of children using data mining techniques using 2011 EDHS datasets. Both are used hybrid methodology of Knowledge Discovery Process by using Weka tools using J48 decision tree and Naïve Bayes and APRT rules induction classifiers. The results from this study were encouraging and confirmed that applying data mining techniques could indeed support a predictive model-building task that predicts nutritional status of under-five children in Ethiopia.

Abera[26] 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 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 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 Page 41 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 have considered in the area.

Another study has been conducted by Amanuel [30] using data mining techniques to predict household health seeking patterns using BRHP dataset. The researcher aim was to develop a model that identifies risk factors and patterns of household health seeking behavior at Butajira district. He used 60,446 records for experiments with implementation of J48 decision tree techniques. The finding of the researcher indicated that with an accurate rate of 89.9017%, predicting household health seeking pattern through data mining techniques is possible.

Taddesse[31] 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 developed as experimental analysis to outperform some of the J48 parameters. The models built allow as more flexible with our output and more powerful weapons in our data mining. As the investigator can see from experiment, 90.3% predictive accuracy 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 predicted by the application of classification technique (J48 decision tree algorithm) given the limitation of this study.

In all the researches done, scholars tried to search a new knowledge that may help for business objectives in relation for under five children mortality as an information source on the base of different epidemiological methods like SPSS, EPI-Info, and STATA etc. Using such tools become inefficient to detect unanticipated interesting patterns from

voluminous data [82]. In addition, most of the studies have been stay at a quantitative feature studies with vital information, without considering various bias effects of the data. For utilization of relevant information which is hidden in the data, it is obvious that one need to be engaged in information computational management (data mining technology) since it is efficient to find unrecognized new knowledge and can mine the knowledge rules automatically from the content of data. Therefore, the researcher inspired to prepare predictive model based on that predict under five children mortality pattern.

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

Chapter Three

Methodology

3.1. Overview

Different authors tend to use the terms process model, life cycle or methodology to refer to the same thing. This has led to some confusion in the field.

A process model is the set of tasks performed to develop a particular element, as well as the elements that are produce in each task (outputs) and the elements that are necessary to do a task (inputs). The goal of a process model is to make the process repeatable, manageable and measurable (to be able to get metrics)[47]. Methodology can be define as the instance of a process model that lists tasks, inputs and outputs and specifies how to do the tasks. Tasks performed using techniques that stipulate how they should have to done. After selecting a technique to do the specified tasks, tools can be used to improve task performance[47].

Finally, the life cycle determines the order in which each activity done. A life cycle model is the description of the different ways of developing a project[47]. From the viewpoint of the above definitions, we can define data mining and knowledge discovering process based on those terms.

3.2. Research Design

The main objective of this study was to identify the pattern that hidden in the dataset of 2016EDHS survey. In order to meet the stated aim, the researcher has used Hybrid methodology of KDD (Knowledge Discovery Process) to build a predictive model using data mining techniques. Hybrid process model selected since it combines best features of CRISP-DM and KDP methodology to identify and describe several explicit feedback,

loops which are helpful in attaining the research objectives. This is because of the academic research community more benefits and interested on the research-oriented steps provided [46].

As Fayyad et al. [44] defined that KDP is nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Thus, using the six steps of the hybrid methodology (i.e. problem domain understanding, data understanding, data preparation, data mining, evaluation and use of the discovered knowledge) the researcher predict the pattern of under-five mortality in Ethiopia based on the EDHS 2016 dataset. One of the key features of this model is iterative and interactive nature of the model so any change triggered in one of the steps the feedback loops revised the process. There are several activities and processes, which performed in each of these steps. The following is the description of the activities performed in each step.

3.3. Understanding the Problem Domain

Understanding the problem area is the basic one, which helps to define the problem and to determine the research goal so that it provide a solution for the current problem. Understanding involves the understanding of the terminology, preparation of a description and restriction of the problem.

Demographic health Survey (DHS) first introduced in 1984 by United States Agency for International development (USAID). Therefore, from the establishment up to now 90 countries benefited from this program. DHS program has collected, analyzed, and disseminated accurate and representative data on population, health, HIV, and nutrition through more than 300 surveys in over 90 countries. The DHS surveys typically conducted every five years and usually based on a representative sample size at national, regional and urban-rural residence type.

EDHS 2016 conducted under supervision of Federal Ministry of Health (FMOH) and implemented by CSA. The testing of the blood samples for HIV status handled by the Ethiopia Health and Nutrition Research Institute (EHNRI). ICF International provided

technical assistance as well as funding to the project through the MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide. Funding for the EDHS study, provided by the government of Ethiopia and various international donor organizations and governments. The United States Agency for International Development (USAID), the HIV/ AIDS Prevention and Control Office (HAPCO), the United Nations Population Fund (UNFPA), the United Nations Children is Fund (UNICEF), the United Kingdom Department for International Development (DFID), and the United States Centers for Disease Control and Prevention (CDC).

Child mortality is one of the main concerns in the world so different strategy designed and implemented to reduce mortality. In this research, the researcher has planned to discuss with domain experts. In addition, review different reading material like policy documents for example reviewing Ethiopian health policy has an impacts in understanding the problem domain and different reports which included child health statistics abstracts which is organized by ministry of health; books, journals articles and different related documents that wrote on children health care and data mining techniques in child mortality. Based on this the researcher has a capability to define the problem and determine the objectives.

For this particular study child health data extracted from the women reproductive questioners from EDHS 2016 used. Followed this, the researcher translates the general goal into data mining objectives in the next steps.

3.4. Understanding the Data

EDHS 2016 dataset used to build the child mortality predictive model. EDHS 2016 data was collect by Federal Ministry of Health and CSA in collaboration with different non-governmental organization (NGOs). The collected data contains detailed information about respondents and related to the indicators of under-five children, fertility, family planning methods, child feeding practice, child mortality, nutritional status of mother and children, adult mortality, maternal mortality, utilization of maternal and child health

services, knowledge of HIV/AIDS and prevalence of HIV/AIDS from 16,650 households, 15,683 female and 12,688 male respondents. This survey conducted at national level in nine regional states and the two (Addis Ababa and Dire Dawa) city administrative regions.

Central Statistics Authority (CSA) designed five questioners to collect EDHS data from the field, the household questionnaire, the woman questionnaire, the man's questionnaire, the biomarker questionnaire and the health facility questionnaire. Each of the questioners collect basic demographic information for each person interviews (households), like age, sex, marital status, education, birth history, childhood mortality, adult mortality, maternal mortality, vaccination, childhood illnesses, wealth index, ethnicity, occupation, exposure, place of delivery and relationship to the head of the household, etc.

3.5. Data Preparation

After well understood the problem domain and the data of the projects, the third steps that are need for all activities are preparing the final datasets. Preparing includes selecting, cleaning, deriving, integrating, and formatting data in order to apply specific data mining tasks. For this study, the data which EDHS 2016 dataset taken from Demographic Health Survey (DHS) database. Therefore, based on the proposed objectives of the study the women individual questionnaires were select to apply selected data mining techniques. Then the next task applied on selecting the attribute and datasets, handling missing value, data integration, and data transformation on the specified datasets.

In addition, the data converted from one format to another format to make convenient to WEKA tools. On this study the statistical tools SPSS 20, Excel 2010 are used to solve data related problem like missing value, noises, and exporting data from existing format SPSS 20 to WEKA 3.8.2 understandable format is called Comma Separated Value (.csv) and Attribute-Relation File Format (.arff) were done.

3.6. Building the Data Mining Model

Building a model is one of the major aims of data mining. Model, which helps you to understand and discover the existing data and future data. The outputs provided by modeling algorithms should be evaluate and interpreted to make sure the resulting model is good enough. Thus, trying to use the result of data mining without any validation it might be very risky and directing us to major problems. The fact that knowledge discovery process is not a separate step rather interlinked with data mining process.

There is a number of analysis tools to discover data or/and knowledge from different views points and summarized the information. Weka is one of the data mining tools. In this study WEKA version 3.8.2 has been use for building the model and visualized the results. Weka has many techniques to use [83]. In addition to this, selecting of the appropriate criteria is the main problem on the data mining so to overcome this problem the researcher where build a classification model using commonly applied techniques Random Forest Tree and PART decision rule classifiers.

3.7. WEKA Tools Selection

Today the world changed in rapid ways in that matter the technologies changed day to day to cope the gaps of the change and facilitate those gaps based on human interest. Internet is one of the technology rapid changed. Due to this, the volume of stored data increased and it needs more storage to handle. Beside on the storage, it needs big data analysis; for this Data mining one of the solutions so far used.

Data mining is mining of the hidden information from the data by using different data mining software. Now there are number of data mining software were developed. Weka is the first in datamining subject area and there is more in this area; some of them are (R-software, Orange, pythons, PSP, and so on.)

This software has own capacities and limitations in the implementation area.

WEKA is the product of the University of Waikato (New Zealand) and the first modern form implemented in 1997. The name stands for Waikato Environment for Knowledge

Analysis. It uses the GNU General Public License (GPL). The figure of Weka is show in the figure 3.1. The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (like tables and curves) [84]. It runs on almost any platform and it is test 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 [15].

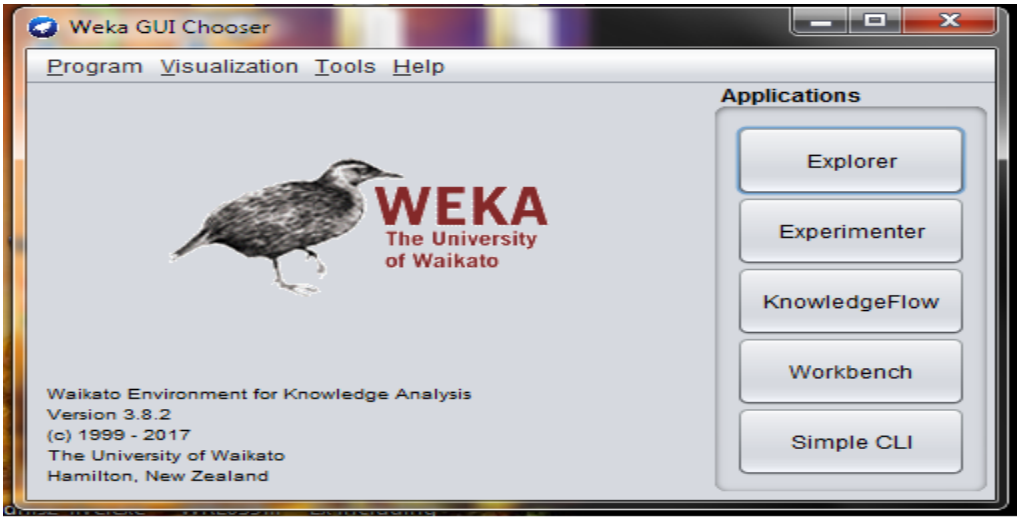


Figure 3. 1: WEKA GUI application main window

A number of DM methods implemented and experimented in the WEKA software. Some of them were based on trees based like the Random Forest, 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 [85].

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 pattern 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 called classifiers, and in the interactive WEKA interface you can 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[15].

Implementations of actual learning schemes are the most valuable resource that WEKA provides. However, 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[84]. The researcher showed how different filters could use, list the filtering algorithms, and describe their parameters in the prediction of the pattern of under-five mortality in Ethiopia.

The data 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 commas separate the attribute values for each instance. 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[15].

3.8. Random Forest Classifier

Random forests are a representative data mining of classification method that uses many trees for the purpose. Random forests are known to be robust for real world datasets that may not have enough information as well as may have missing and some error data[86]. Therefore, random Forest is a computationally efficient technique that can operate quickly over large datasets. It has been used in many recent research projects and real-world applications in diverse domains[87, 88] and it is suggested by Breiman are based on BAGGING[88] use many decision trees with some random selection of attributes

to split each node in the tree, and do unpruned. In other words, random forests use bootstrap method in sampling to generate a training set, and the training set is used to build a tree, and since bootstrap method uses sampling with replacement, each training set can have some duplicate instances and could compensate the insufficiency of data to train somewhat. The random Forest technical algorithm adopted[89, 90] here as follows.

Steps:

1. Randomly select an observation from the original data
2. "Write it down"
3. "Put it back" (i.e. any observation selected more than once because selection conducted always with replacements)
4. Repeat steps 1-3 N times; N is the number of observations in the original sample

FINAL RESULT: One "sample tree" with N observations

5. Draw each of a sample Tree
6. Fit a large, unpruned tree to this sample tree

-At each split in the tree consider only k randomly selected variables instead of all of them

7. Repeat Steps 5- 6 M times
8. The final prediction is determined via majority vote or via a cutoff probability

Predict a New Record: run the record down each tree, each time computing a prediction

Final Prediction for a New Record:

Total "Class1" Predictions for the New Observation: X=300

Total "Class2" Predictions for the New Observation: Y=200

Final Prediction for the New Observation: "Class1"

Predicted Probability for Class1 for the New Record:

(X/(X+Y))*100 ===>300 /500 = 60%

After sampling some conventional decision tree algorithms like C4.5 or CART applied, but without pruning. When random selection of attributes to split each node is applied, some predefined number limits the number of candidate attributes for split, say user may give

R. R, or default value used. Default R-value is the first integer less than (log2A+1) and the half and double of the number recommended for further search. So, depending on which number is used, the degree of randomness in tree generation is affected[91].

The other factor that affects the accuracy of random forests is the number of decision tree. Nowadays, Random Forest is a method of ensemble learning widely used in the literature and applied fields. However, the associate literature provides few or no directions about how many trees used to compose a Random Forest. In general, the user sets the number of trees in a trial and error basis[87]. Sometimes when s/he increases the number of trees, in fact, only more computational power spent, for almost no performance gain.

Random Forests are efficient, multi-class, and able to handle large attribute space, they have been widely used in several domains such as real-time face recognition, bioinformatics, and there are some recent researches in medical domain, as well as medical image segmentation[87].

A tracking algorithm using adaptive random forests for real-time face tracking proposed, and the approach was equally applicable to tracking any moving object. Random Forests, support vector machines, and artificial neural network models developed to diagnose acute appendicitis. Random Forests are used into detect curvilinear structure in mammograms, and to decide whether it is normal or abnormal. It is introduced an efficient keyword based medical image retrieval method using image classification with Random Forests. A novel algorithm for the efficient classification of X-ray images to enhance the accuracy and performance using Random Forests with Local Binary Patterns presented. An enhancement of the Random Forests to segment 3D objects in different 3D medical imaging modalities proposed in [33]. Random Forests evaluated on the problem of automatic myocardial tissue delineation in real-time 3D echocardiography. A new algorithm presented for the automatic segmentation and classification of brain tissue from 3D MR Scans. An automatic 3D Random Forests method which is applied to segment

the fetal femur in 3D ultrasound and a weighted voting mechanism is proposed to generate the probabilistic class label is developed[87].

3.9. Decision Lists (PART-Rule) Classifiers

Classification is a supervised procedure that learns to classify new instances based on the knowledge learnt from a previously classified training set of instances. It takes a set of data already divided into predefined groups and searches for patterns in the data that differentiate those groups supervised learning, pattern recognition and prediction. Typical Classification Algorithms Are Decision trees, rule-based induction, neural networks, genetic algorithms and Bayesian networks. Rule based classification algorithm also known as separate-and-conquer method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover[92].

PART means partial decision trees to generate the decision list that is shown in the output, but only this final list is what is used to make classifications[15]. It is a separate-and-conquer rule learner proposed by EIBE and Witten[15, 93]. The algorithm producing sets of rules called decision lists, which ordered set of rules. A new data compared to each rule in the list in turn, and the item assigned the category of the first matching rule (a default applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning[92, 94].

To make more clear let compare PART with other Rule make Algorithms (like RIPPER (Repeated Incremental Pruning to Produce Error Reduction) and RIDOR (RIPPLE-DOWN Rule)).

The RIPPER algorithm is a direct method, i.e. RIPPER extracts the rules directly from the data. The algorithm progresses through four phases: i) growth, ii) pruning, iii) optimization, IV) Selection. In the growth phase, one rule generated by greedily adding attributes to the rule until the rule meets stopping criteria. In the following prune phase, each rule incrementally pruned, allowing the pruning of any final sequence of the attributes, until a pruning metric fulfilled. The optimization stage of each generated rule further optimized by a) greedily adding attributes to the original rule and b) by independently growing a new rule undergoing a growth and pruning phase, as described above. Finally, in the selection phase, the best rules kept and the other rules deleted from the model. RIDOR is also a direct method, first generating a default rule to the default rule with the least error rate. The "best" exceptions for each exception are generated and iterated until pure. Thus, a tree-like expansion of exceptions generated. The exceptions are a set of rules that predict classes other than the default. PART is an indirect method for rule generation. PART generates a pruned decision tree using the C4.5 statistical classifier in each iteration [95]. From the best tree, the leaves translated into rules.

PART (Partial Decision Tree)[15] is an indirect technique for constructing classification rules. It employs partial decision tree to generate the individual rules and the tree induced with C4.5 classifiers. After tree generation, rules derived directly from the partial tree starting with the deepest leaf node, in combination with every node along the path towards the root. Then, the partial decision tree removed. The algorithm for generating partial decision tree and PARTRule algorithm for classification across multiple database relations, which is adopt from papers as shown as follows[15, 96, 97].

Algorithm: PARRule (D, Rt)

Input: A relational database D with a target Relation Rt

Output: A Partial decision tree N for predicting class labels of target tuples

Procedure: N = empty; Tree node initially empty

A = empty; Attribute for storing max. Information gain

If |Rt| < MIN_SUP.

Then return Evaluate all attributes in any active relation or Relations joinable with active relation using Information gain

A = attribute with max information gain

If (info_gain (A) < MIN_INFO_GAIN) then return

N = A; set relation of A to active

Do Divide Rt into subsets R′ according to A and add those nodes as children of node N

While (there are subsets that have not been expanded and all the subsets expanded are so far are leaves)

If all the subsets expanded are leaves and estimated Error for sub tree >= 0.5 Undo expansions into subsets and make node a leaf

For each relation R that is set active by this function Set R′ inactive

Return N

End

3.10. Performance Evaluation for Predictive Model

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

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 EDHS 2016 dataset.

Predictive models evaluated in terms of correctness, often referred to as performance, and applicability. The performance measures almost geared towards the evaluation of an instance of a model type, and are 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 [73].

Once a predictive model developed using, the under-five mortality EDHS 2016 dataset, the model should 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 used.

3.11. Confusion Matrix

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 implement to confirm the model performance evaluation for the results of the predicted model of the under-five mortality in Ethiopia. Given M classes, a confusion matrix is a table of at

least size M by M. An entry, CMi,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 [98]. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry CM1, 1 to entry CMm,m, with the rest of the entries being close to zero [13, 99].

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 [100].

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Table 3. 1: Confusion Matrix Adopted from Jiawei Han 2011[13]

As shown in table3.1, 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 which is adopted from Jiawei Han [15, 100].

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

The True Negative Rate of a classifier estimated by dividing the incorrectly classified negatives by the total negatives count which is adopted from Jiawei Han [15].

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

The Accuracy of a classifier projected by dividing the total correctly classified positives and negatives instances by the total number of samples, which adopted from Jiawei Han[15].

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

Precision calculated by dividing correctly classified instances by the total number of correctly and incorrectly classified samples, which adopted from Jiawei Han[15].

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

F-Measure calculated as the harmonic mean of recall and precision, which adopted from Jiawei Han[15].

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

3.12. Receiver Operating Characteristics (ROC) Analysis

Another useful method for evaluating classification models is Receiver Operating Characteristics (ROC) analysis. A large number of intelligent medical systems (including medical expert systems, neural networks, classifiers, knowledge discovery and data mining systems) showed great progress and they are being develop. Practically to aid clinician

and to improve patient care in areas such as diagnosis, prognosis, decision support and screening. To test which classifier is highly significant for a given subject is determined by ROC analysis and it is becoming widely used tool in medical tests evaluation [42, 101].

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified [11, 102]. 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)[103]. 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 achieved. This threshold can then be use as a post-processing parameter for achieving the desired performance with respect to the error rates.

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[73] to the case of three classes for which a predictive model returns a probability distribution. The AUC measure has been extend to the volume under the surface (VUS). Therefore, the researcher tried to implement ROC analysis to evaluate the predicted model constructed by the Random Forest algorithm and PART classifier in order to get the best-fit model for the domain area of health sector specially to predict the pattern of under-five mortality in Ethiopia.

Chapter Four

Data preprocessing and preparation

4.1. Overview and Data Mining Goals

This chapter provides interesting features for business understanding and data preprocessing of the EDHS 2016 dataset that helps to measure accurately the under-five mortality.

Under-five mortality has long been used as indicator of the level of socioeconomic development of a nation[22, 104]. 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[105, 106].

To improve the health care planning and strategies in Ethiopia in focus of under-five child health, in this research critical assessment will be conduct in detail to select constructive under-five child mortality predictive model. Then before starting the actual data mining task, we should be able to clearly define our problem and have a good understanding of our data to use for the datamining task.

As Two Crows Corporation (1999) defined, "identifying the goal of the data mining process is a prerequisite to discover knowledge form the database"[65]. Based on this, in chapter one at section 1.3 clearly define the problem of the study and understand the data mining tasks. Thus to remind once again, the main aim and outcomes of this research is to identify the determinant factors and find interesting and meaningful patterns and relationships in child mortality using EDHS datasets. So, it provided that meaningful relationship among attributes to establish a prevention programs for child mortality and to have a better understanding of the nature of child mortality in Ethiopia and thus could develop strategic solution to avoid the most determinate factors of child mortality. Now, here this is not discussed again rather the remaining five basic literature steps of hybrid

methodology[38] discussed. Because the success of all the other steps depending on what extent the problem is clear defined and the dataset, selection for mining related to the business area.

A hybrid data mining method is adapted which guide the researcher to apply data selection and pre-processing in order to extract useful information from the dataset using DM technique. As this is the very important step of the knowledge discovery process, the researcher has made a great deal of effort to understand the business domain, which enables to select the major variables in achieving the objectives of the study. After selecting the major variables that affect under-five child health with the assistance of the domain expert and by reviewing related literature, the researcher has employed data preprocessing task. Data pre-processing is an essential step in the process of preparing dataset that is appropriate for mining. The purpose of data pre-processing is to clean the noisy data, extract and merge the data from different sources, and then transform and convert the data into a proper format[102]. It is an important step in data mining, because quality decisions must base on quality data.

This section briefly deals about the data source and selection attribute or feature selection, data cleaning and data transformation and aggregation, integration and formatting of the data employed in this study.

4.1.1. Data Understanding and Data Source Description

After setting up the problem and a rough plan for its solution, the researcher proceeded with the central item in data mining process - data. There are several things to learn about the data before the actual application of data mining techniques.

4.1.1.1. Under-five Mortality Based on EDHS 2016 Dataset

Using the right data for data mining task is one of the primary keys for successful data mining [65]. The data source for this study obtained from the organization's internet website address (i.e. www.dhsprograms.com) after getting consent from the organization. The EDHS 2016 conducted by the Central Statistical Agency (CSA) under the umbrella of the Ministry of Health. It is the Fourth Demographic and Health Survey (DHS) conducted in Ethiopia, under the worldwide MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide. The survey interviewed a nationally representative population in about 16,650 households. Out of these households, a nationally representative sample of 15,683 women of age 15–49 and 12,688 men of age 15–59 interviewed. This represents a response rate of 95% for women. The data were collected on key indicators relating to family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women's empowerment, and knowledge of HIV/AIDS [16].

The primary objectives of the 2016 EDHS were to provide up-to-date information for planning, policy formulation, monitoring, and evaluation of population and health programs in the country.

Since the goal of this study was to predict patterns of under-five child mortality, the dataset used from women is of reproductive questionnaires. The Woman's data used to collect information from all women age 15-49. These women asked questions on the following topics:

- Background characteristics such as age, education and media exposure
- Birth history and childhood mortality

- Adult mortality, including maternal mortality
- Knowledge and use of family planning methods
- Fertility preferences
- Antenatal, delivery and postnatal care
- Breastfeeding and infant feeding practices
- Vaccinations and childhood illnesses
- Women's work
- Husband's background characteristics
- STIs (Sexual Transmitted Infection)

After the initial data collection, new database created with MS Excel formats. MS Excel used for preparing the dataset into a form acceptable by the selected data mining software Weka. There are 109 attributes which are taken from the three different forms of questionnaire (called Women's, Child (Kid's) and Birth Records), to take those attributes from these three datasets, margining variables functions technology were applied on SPSS software.

4.1.2. Variable Selection on Maternal Healthcare Service Utilization

Variable selection on under-five child mortality deciding on the data that will be used for the analysis is based on several criteria, including its relevance to the data mining goals, as well as quality and technical constraints such as limits on data volume or data types[16]. Therefore, in this research the attributes selected with the help of domain expert and extensive literature review. Because taking all the variables in the database we have, fed

them to the datamining tool and find, which are the best predictors, may not work very well. Some of the reasons are the following:

The first reason is that the time it takes to build a model increases with the number of variables increases.

The Second reason is that blindly including extraneous columns can lead to incorrect models[11].

Although in principle, some data mining 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 knowledge of the problem domain helps to make these selections correctly[65].

Thus, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts in order to simplify the task of modeling. Accordingly, the following main variables selected from the women's individual record dataset. Therefore, Mother's age at birth, Place of Residence, Region, Household Wealth Index, Religion, Highest educational level, Husbands education level, Total children ever born, Frequency of listening to radio, Frequency of reading newspaper, Frequency of watching television, Husbands/partners occupation, Mother's occupation, Number of antenatal care visit, Delivery care and Postnatal checkup are the major variables that determine under-five children mortality. These variables provide the socioeconomic and demographic information for each respondent.

Attributes with no variation in their value throughout the dataset and attributes which is serve for assigning sequence number for the records were all exclude. Attributes, which have missing value for more than 95% of instances, also cancelled. Out of the very brief 109 lists of attributes in the selected dataset, 41 attributes eliminated because they have

missing values in more than 90% of records. In addition, the other twelve attribute again eliminated since their value is constant throughout the database or they used simply to assign sequences to the records. At this point, the number of attributes diminished to 56. Since Random Forest tree induction algorithms capability of predicting the pattern from large dataset, so to have good accuracy classification it depends on the data we have at our hands; missing values considerably reduce their performance. When the number of missing values becomes higher, the pattern to generate by the classification algorithm reduces continuously. After this further attribute elimination, the number of attributes taken for analysis was 53. Domain expert help obtained in analyzing the importance of the attributes to the datamining goal and fixing the threshold for missing values. In addition to domain expert feedback, the researcher was used the information gain result for the selection.

4.1.3. Description of the Selected Attributes

The description of the selected attributes with their data type, values, and percentage of missing values are depict in the following section. The final selected attributes were prepared and pre-processed before developing the model.

No	Attributes	Description	Values	Data type	Count Missing
1	Mother's age	Current age of mother	Mother age in five year interval group (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49)	Categorical	0(0.0%)

2	Place of residence	The mother/child living place	Residence (Rural or Urban)	Category	0(0.0%)
3	Women's education level	The highest level of education attained by women	No education, Primary, Secondary, Higher.	Category	0(0.0%)
4	Household Wealth Index	The living standard of the household	Poorest, Poorer, Middle, Richer, Richest	Category	0(0.0%)
5	Religion	Religion of the mother	Orthodox, Muslim, Protestant, Catholic, Traditional, Others	Categorical	0(0.0%)
6	Husbands education level	Education status of a partner	No education, Primary, Secondary, Higher.	Categorical	535 (7.76%)
7	Total children ever born	Number of children of the women deliver in life of her.	1 child, 2 or 3 child, 4 or 5 child, more than 6 child	Numeric	0(0.0%)
8	Frequency reading newspaper	How often do a women reading newspaper to radio	Not at all, Less than once a week, At least once a week	Categorical	0(0.00%)
9	Frequency of listening radio	How often do a women listening radio	Not at all, Less than once a week, At least once a week	Categorical	0(0.00%)
10	Frequency of watching Television	How often do a women watching Television	Not at all, Less than once a week, At least once a week	Categorical	0(0.00%)

11	Husbands /partners occupation	Job of a partner	Agricultural-employee, clinical, professional/technical/managerial, Sales, Service, Did not work, Skilled Manual, Unskilled manual.	Categorical	557 (8.08%)
12	Mother's occupation	Respondents occupation	Agricultural-employee, clinical, professional/technical/managerial, Sales, Service, Did not work, Others, Skilled Manual, Unskilled manual.	Categorical	0(0.0%)
13	Number of Antenatal Visit	Number of Maternal healthcare service received during pregnancy	One visit (1), two visit (2)……., not antenatal visit	Categorical	18(0.26%)

Table 4. 1: Description of the Selected Attribute from EDHS 2016 dataset

4.1.4. Statistical Summary of the Selected Attributes

The summary of each of the selected attributes used for model building statistically described in detail in Table4.2. This statistical summary of the attributes is helpful for understanding of the data set for DM model building phase.

No.	Variables	Frequency	Percen
1.	**Respondent's Age**		
	15-19	324	4.70%
	20-24	1412	20.48%
	25-29	1933	28.04%
	30-34	1494	21.67%
	35-39	1106	16.05%
	40-44	469	6.80%
	45-49	155	2.25%
	Missing	0	0
	Total	6893	100.0
2.	**Place of residence**		

	Rural	5474	79.41%
	Urban	1419	20.59%
	Missing	0	0
	Total	6893	100.0
3.	**Women's highest educational level**		
	Higher	289	4.19%
	No education	4205	61.00%
	Primary	1854	26.90%
	Secondary	545	7.91%
	Missing	0	0.0
	Total	6893	100.0
4.	**Household Wealth Index**		
	Middle	986	14.30%
	Poorer	1140	16.54%
	Poorest	2342	33.98%
	Richer	877	12.72%
	Richest	1548	22.46%
	Missing	0	0.0
	Total	6893	100.0
5.	**Religion**		
	Catholic	46	0.67%
	Muslim	3183	46.18%
	Orthodox	2260	32.79%
	Other	49	0.71%
	Protestant	1291	18.73%
	Traditional	64	0.93%
	Missing	0	0.0
	Total	6893	100.0
6.	**Husbands education level**		
	Higher	537	7.79%
	No education	3016	43.75%
	Primary	2098	30.44%
	Secondary	707	10.26%
	Missing	535	7.76%
	Total	6893	100.0
7.	**Total children ever born**		
	1 child	1338	19.41%
	2 child	1160	16.83%
	3 child	986	14.30%
	4 child	840	12.19%
	5 child	749	10.87%
	6+ child	2820	26.40%
	Missing	0	0.0
	Total	6893	100.0

8	**Frequency reading newspaper**		
	At least once a week	130	1.89%
	Less than once a week	434	6.30%
	Not at all	6329	91.82%
	Missing	0	0.0
	Total	6893	100.0
9	**Frequency listening radio**		
	At least once a week	879	12.75%
	Less than once a week	871	12.64%
	Not at all	5143	74.61%
	Missing	0	0.0
	Total	6893	100.0
10	**Frequency of watching Television**		
	At least once a week	1002	14.54%
	Less than once a week	582	8.44%
	Not at all	5309	77.02%
	Missing	0	0.0
	Total	6893	100.0
11	**Husbands /partners occupation**		
	Agricultural – employee	3419	49.60%
	Clerical	40	0.58%
	Did not work	646	9.37%
	Others	285	4.13%
	Professional/technical/managerial	459	6.66%
	Sales	508	7.37%
	Services	253	3.67%
	Skilled manual	473	6.86%
	Unskilled manual	253	3.67%
	Missing	557	8.08%
	Total	6893	100.0
12	**Mother's occupation**		
	Agricultural – employee	1391	20.18%
	Clerical	38	0.55%
	Not working	3920	56.87%
	Others	132	1.91%
	Professional/technical/managerial	133	1.93%
	Sales	787	11.42%
	Services	147	2.13%
	Skilled manual	245	3.55%
	Unskilled manual	100	1.45%
	Missing	0	0.0
	Total	6893	100.0
13	**Number of Antenatal Visit**		

	1 Visit	331	4.80%
	2 Visit	547	7.94%
	3 Visit	1140	16.54%
	4 Visit	1090	15.81%
	5 Visit	593	8.60%
	6 Visit	380	5.51%
	7+ Visit	424	6.15%
	No antenatal visits	2370	34.38%
	Missing	18	0.26%
	Total	6893	100.0

Table 4. 2: Statistical Summary of the Selected Attributes

4.2. Data Preprocessing

Much of the raw data contained in databases is un-preprocessed, incomplete, and noisy [102, 107]. For example, if the databases contain:

- Missing values
- Outliers
- Data in a form not suitable for data mining models

To be useful for data mining purposes, the databases need to undergo pre-processing, in the form of data cleaning and data transformation. Data mining often deals with data that hasn't been looked at for years, so that much of the data contains field values that have expired, are no longer relevant, or are simply missing [102].

4.2.1. Handling Missing Value

Missing data is a problem that continues to plague data analysis methods. Even as our analysis methods gain sophistication, we continue to encounter missing values in fields, especially in databases with a large number of fields. The absence of information is rarely beneficial[102]. All things being equal, more data is always better. Therefore, we should think carefully about how we handle the thorny issue of missing data [102].

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[102]. Therefore, data analysts have turned to methods that would replace the missing value with a value substituted according to various criteria.

- Replace the missing value with some constant, specified by the analyst.
- Replace the missing value with the field mean (for numerical variables) or the mode (for nominal variables) or median (for ordinal variables).
- Replace the missing values with a value generated at random from the variable distribution observed.

All attributes with missing values among the selected attributes in this research are nominal. Therefore, the second approach used for handling missing value in a dataset as follows:

No.	Attribute	Frequency of missing value	Percentage of missing values	Replaced value
1	Husbands/Partners occupation	557	8.08	1 (Agriculture –employee)
2	Husbands education level	535	7.76%	0(No education)
3	Number of Antenatal Visit	18	0.26%	3 (Visit)

Table 4. 3: Attributes with missing values replaced by mode

4.2.2. Data Quality Assessment and data cleaning

A data quality assessment identifies characteristics of the data that will affect the model quality. Data cleaning is the process of examining data and determining the existence of incorrect characters and mis-transmitted information[65]. In this phase, the researcher attempted to ensure not only the correctness and consistency of values but also that all the data measured in consistent way. One of the data mining software used in this research project, Weka, forces the use of clean data. Weka would not open a data file unless it is clean and in required format. This data mining software has a facility for data analysis. An option named 'Analysis' helps the user to clean the data by pointing out the type and the position of error in the dataset. Weka considers any attribute value, which is not compatible with its requirements as error. The researcher used this option to clean the entire target data selected from the women reproductive dataset.

4.2.3. Data Transformation, Reduction and Reformatting

This task includes constructive data preparation operations such as the production of derived attributes, creating new records or transformed values for existing attributes, consolidating and amalgamating records and summarizing fields.

The data may also need to be transform into forms appropriate for mining. The process of data transformation might include smoothing (e.g. using bin means to replace data errors), Normalization, where the attribute data are scaled so as to fall within a small specified range (scaling the data inside a fixed range), and Attribute construction, where new attributes are constructed and added from the given set of attributes to help the mining process [11].

The data needed to be reduce in order to make the analysis process manageable and cost efficient. Data reduction techniques include a data discretization technique, which use to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values, dimension reduction (irrelevant or redundant attributes are removed), and data compression (data is encoded to reduce the size, numerously reduction (models or samples are used instead of the actual data) [11].

From the original dataset, the attribute of number of living children in the house is a continuous variable on which discretization performed to convert into three distinct values as shown in table below:

Those intervals chosen according to Ethiopian demographic health survey reports. The report says total fertility rate in Ethiopia is 4.5 children per women. Therefore, this means one woman can have born on average up to 5 children in her life. Therefore, the three distinct values chosen to meet those criteria by equally width of discretized.

Number of children	Represented value
(-inf-5.333]	1-5
(5.333-9.667]	6-10
(9.667-inf)	11-14

Table 4. 4: A discretized number of living children

The attribute birth order of child in the house is a continuous variable on which discretization performed to convert into four distinct values as shown in Table 4.5 below:

The discretized value is represented in to four similar range of values to simplicity and easily interpretation so that the any (the upper and lower) decimal points are not rounded.

Birth order of child	Represented value
(-inf-4.25]	1-4
(4.25-7.5]	5-7
(7.5-10.75]	8-10
(10.75-inf)	11-14

Table 4. 5: A discretized Birth order of children

In addition to this the value of number of antenatal visit reformatted based on 2016 WHO guideline and currently used by ministry of health by adapted from WHO guidelines. Therefore, the EDHS 2016 datasets contain the pregnant women were visit the health facility up to 14 times (the min 1 and max 14) throughout the period of her pregnancy. Therefore, based on the guideline of federal ministry of health one pregnant mother should visit at least four times as standard and the rest visit consider as additional visit sometimes it considered as clinical visit. Thus, 1-4 visit grouped separately and the rests (5+) grouped like as additional visits. Based on this the number of antenatal visit categorized as follows:

Number of antenatal Visit	Represented value
1	One Visit
2	Two Visit
3	Three Visit
4	Four Visit
5	Five and More Visit

Table 4. 6: A reformatted the antenatal Visit of pregnant women's

From the original dataset the Husband and Women's Occupation level the respondent's answers have nine range of answers however to be convenient for analysis purpose the nine response grouped in to five responses based on the international standard classification occupation definition(ISCO)[108]. Therefore, the categorization looks like as follows:

Husband/Women's Occupation level	Represented value
Agricultural-Employee + Skilled Manual	AgreEmp
Not Working	Not Working
Services +Clerical + Unskilled Manual+ Sales	Services
Professional/Technical/Managerial	Professional
Others	Others

Table 4. 7: A reformatted (merged) the husband and women's working occupation based on ISCO

Other attributes which need transformation are Frequency of reading news, Frequency of listening to radio, and Frequency of watching television each of them has three values (Not at all, Less than once a week, and At least once a week).

The three attributes combined and a new attribute named Media Exposure created. If a woman has access at least any of the three media once in a week, then it considered she has a media exposure, otherwise not.

Not only had this but also, exposure to drug (ExposureDrug) is the new attribute created from the three variables (the exposure of chewing chat, the exposure of drink alcohol, and the exposure of smoking) which has two answers of value (Yes or No). Therefore, if the women have tested at least one of the three drugs, then it considered she has a drug exposure, otherwise not.

4.3. Balancing Class Variable

One of challenging things in data preparation was changing the imbalanced datasets in to balanced datasets. The term data-imbalancing is common in data mining activities because the development of the data mining application area leading to increase the size of data. Those increasing of data also unbounded in size and imbalanced nature of data[109, 110]. class balancing means the class distribution of the datasets should not be

differ from 50:50 or 60:40 proportion else, we call it class-imbalancing problems (that means the majority of the class larger than the minority class) [111-113]. Those majority of the class larger means the of prediction percentage tilted to the majority class or the minority class preference always less than compared that from the majority class so this lead us to make unbalanced decision.

In our cases, the EDHS 2016 under-five child mortality datasets "child is Live" is selected as class attribute from 21 attributes. However, this class attribute has balancing problem, which called imbalanced. The proportion of "Child is live" 6750/143 (47:1) which means child who born live 47 times greater than child does who die. Therefore, this shows that there is balancing problems thus the prediction of child mortality will be misclassifying in to as all is live born.

There are a number of techniques recommend to correct this imbalancing problems and most of literatures suggested that under-sampling and oversampling[109, 114, 115]. Synthetic Minority Oversampling Technique (henceforth SMOTE) shown good performance for child mortality datasets so it is applied on Weka software whereas 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 the minority class increases. One of the drawbacks of Oversampling (SMOTE) techniques is over-fitting problem when we increase the number of minority class on the datasets but to reduce this over-fitting problem the researcher were used randomizing methodology to distribute the number of minority class throughout the datasets[109, 116].

However, the main question is how many time the number of instance to be increase to get the best performing randomized datasets because, as the number of instance differ the performance of the datasets also differ.

10-Fold Cross validation			
Number of Instances Increases	Algorithm		
	Random Forest	J48	PART
20 Times Increase	97.8979%	96.3597%	96.9340%
30 Times Increase	98.2293%	96.9505%	97.4334%
45 Times Increase	98.4618%	97.3888%	97.8090%

Table 4. 8: A performance of classifiers based 10-fold cross validation after minority class increases

Percentage Split -66%			
Number of Instances Increases	Algorithm		
	Random Forest	J48	PART
20 Times Increase	97.8589 %	95.8251%	96.1399%
30 Times Increase	98.0537%	96.1336%	96.7912%
45 Times Increase	98.4992%	97.0205%	97.5944%

Table 4. 9: A performance of classifiers based Percentage Split-66% after minority class increases

According to above tables, when the number of minority classes increased the more performance of classifiers also increased. That means the best performing classifiers identified when the number of minority classes increased by 45 times from the original minority of class than increased by 20 and 30 times the minority class. Addition to these Random forest algorithms is the best performing classifier than J48 and PART; beside on this, when comparing test options 10-fold cross validation better than percentage splits-66%.

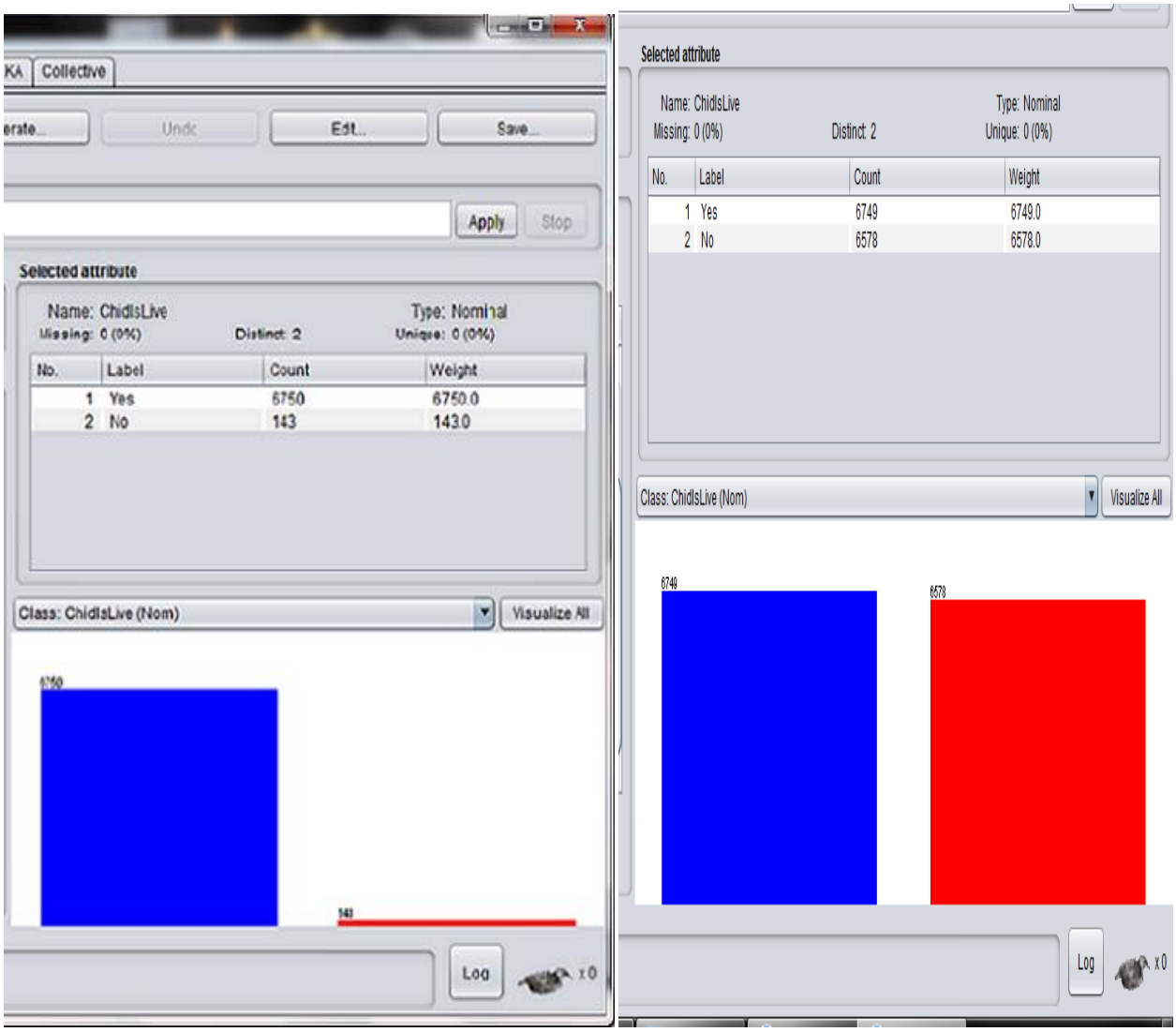


Figure 4. 1: The count of class variable: (a) Original data; (b) Balanced data using SMOTE

Figure 4.1 shows that the number of class attribute status after SMOTE applied to the minority class. Originally, there were 6,750 records in the majority class and only 143 records in the minority class but after applying SMOTE, the minority class increased to 6,578 records and then it randomized to generate new random datasets on a distribution function to control over-fitting problems.

4.4. Final Attribute Selection (Weka attribute selector's classifiers)

To select the final attributes, the researcher used attribute selection method from WEKA packages because fewer attribute is better to classification. There are three type of attribute selection method called wrapper, filter methods and Embedded methods [117, 118]. Each has its own advantages and disadvantages. The advantage of wrapper methods is the capability to remove the irrelevant and redundant attributes and on the results menu it displays the number of sub-attributes from the total list of attributes. The disadvantage is that, its speed is very, very slow.

The advantage of filter methods has the capability of removing the irrelevant attributes and the speed also very fast on the process of selecting the attribute however it is not reduced redundant attribute and generate the results by ranking so the more relevant at the top and the less relevant at bottom[119]. Embedded model proposed to bridge the gap between the filter and wrapper models. First, it incorporates the statistical criteria, as filter model does, to select several candidate features subsets with a given cardinality. Second, it chooses the subset with the highest classification accuracy[117]. Thus, the embedded model usually achieves both comparable accuracy to the wrapper and comparable efficiency to the filter model. The embedded model performs feature selection in the learning time. In other words, it achieves model fitting and feature selection simultaneously[117]

Previously, we have done balancing the class distribution by increasing the minority classes 40 times and now, before conducting selecting the reduced attribute using attribute selection classifiers lets choose the best attribute selection classifiers algorithms based on their performance from 21 attributes by using the most common test option, i.e. cross validation, using training sets and percent (%) split test options.

No	Criteria	Random Forest		
		Cross Validation (10 fold)	using training sets	% split test options (66% Training)
1	No-selection	98.4618%	99.7074%	98.4992%
2	CfssubsetEval.(Fast)	96.6834%	97.6214%	96.6674%
3	GainRatioAttributeEval	97.1486%	98.2292%	97.0647%
4	(InfoGainAttributeEval	98.0266%	99.2571%	97.8371%
5	AttributeSeclectedClassifier(slow)	98.4693% (Gainration)	99.7074% (Gainration)	98.4551% (Gainration)

Table 4. 10: A performance of classifiers on different test options with Random Forest

Before going to generate the classifier comparison using Random forest algorithm, here let us see that the performance of test options. The training sets is the best performing test option than any other two test options and second one is cross validation with 10-folds preferred because it is better than percentage (%) split tests. Therefore, cross evaluation always the first choice by many scholars than any other test option. Because it helps to ensuring each instances used for training and testing an equal number of times (the average of the K-folds) while reducing the variances of an accuracy and it also uses randomness to decide how to split the data sets in to K-folds.

Then, when compared the classifiers based on their performance; attribute selection classifier are the best performing classifiers than other classifiers. Therefore, attribute selection classifiers were good classifiers for this datasets using Random Forest algorithms.

Let's compare the PART with different attribute selection techniques.

	Criteria	PART		
		Cross Validation (10 fold)	using training sets	% split test options (66% Training)
1	No-selection	97.8090%	98.5143%	97.5944%
2	CfssubsetEval.(Fast)	97.5763%	98.1241%	97.1529%
3	GainRatioAttributeEval	97.4188%	97.8840%	97.0205%
4	(InfoGainAttributeEval	97.4638%	98.0566%	97.0647%
5	AttributeSeclectedClassifier(slow)	97.8390% (infoGain)	98.5518% (infoGain)	97.5944% (infoGain)

Table 4. 11: A performance of classifiers on test options with PART Decision lists

Here, you see that it is similar to with the above Table 4.10, the performance of training sets is the best than any other two test options from the classifiers listed in the above Table 4.11 and rests (cross validation and percentage (%) of split tests) ranked seconds and thirds respectively. Moreover, when compare the performance of classifiers the attribute selection classifiers performed higher and it is better to use attribute selection classifiers to select the more related attribute with class attribute.

So based on the attribute selection classifier the attribute evaluator ranked the attributes with the bests performing attributes ranked first and the least performing lasts. Therefore, among the attributes those listed at the bottom of the list the researcher deleted attributes by tested the performance with different attribute combination. Therefore, after removing some of the attributes based on their law performance and the performance of the classifiers were improved than the whole attributes performing and the domain experts also agreed with the removal of those attribute because of deleted attributes are not that much important for outputs analysis.

Finally, after pre-processing the original dataset there are 18 attributes were choose (17 Independent attributes and 1 Dependent attributes) and 13,327 instances from among 18 attributes and then those assumed to be relevant to the target variable, for constructing the model.

The final selected dataset with their description summarized in Table 4.12.

No	Attribute Name	Data Type	Description
1	Womenage	Nominal	15-19=1 , 20-24=2 , 25-29=3 , 30-34=4 , 35-39=5 , 40-44=6 , 45-49=7
2	PlaceRes	Nominal	Rural , Urban
3	WomenEdu	Nominal	No Education , Higher, Primary, Secondary
4	Religion	Nominal	Orthodox, Muslim, Protestants, catholic, Traditional and Other
5	Wealth	Nominal	Poorer, Poorest, Richer, Richest, middle
6	ExposureMedia	Nominal	No media exposure, Has Media Exposure
7	TerminatePreg	Nominal	Yes, No
8	AnyUsedDelayPreg	Nominal	No , Yes Used Calendar, Yes, used Outside Calendar
9	HusbandEdu	Nominal	No Education , Higher, Primary, Secondary
10	HusbandOcup	Nominal	AgreEmp, Not Working, Services, Professional, Others
11	WomenOcup	Nominal	AgreEmp, Not Working, Services, Professional, Others
12	NumANCvisit	Nominal	No antenatal Visit, One visit, Two Visit, Three Visit, Four Visit, and Five and More
13	SizeChild	Nominal	Average, Larger than average, Very large, Smaller than average, Very small, Do not know
14	DrugTakenPreg	Nominal	Yes, No
15	ChildVaccine	Nominal	Yes, No
16	ExposureDrug	Nominal	Has Drug Exposure, No Drug Exposure
17	AnythingOnCord	Nominal	Yes , No
18	ChildIsLive	Nominal	Yes , No

Table 4. 12: Final selected variables with their description

4.5. Choosing the Best Classifiers Algorithms Relative to Datasets

Choosing the best attribute is one of the most important activity in data mining analysis because there is a number of algorithms in Weka but all of them may not relevant or may not fit for the datasets[120]. Therefore, we need to test the value of some of the attribute based on the objective of the project and the performance of the algorithms capacities.

For this study several techniques are compared using the analysis of Imbalancing EDHS 2016 datasets rather than using traditional classification techniques[121].

The EDHS2016 datasets used to build classifiers and test the performance. In this experiment, the performance of the classifiers progressively increases in the datasets of imbalanced class by oversampling (SMOTE) in the minority class of child mortality, to identify to what extent the predictive power of the respective techniques is adversely affected.

To choose the performance criterion and to measure the classifiers effect is the two Weka application used. The first one is Weka GUI experimenter application and the second; Auto-Weka application techniques applied. Therefore, based on their accuracy performance results the classifiers ranked as follows using the two methods.

4.5.1. Automatic Model selection

Weka is providing the new technology to choose the best classifiers that is called Auto-Weka 2.0 for automatic model selection [122] rather than using the only application called experimental application GUI methods.

The main objective of this new Auto-Weka technology were designed due to the interest of Weka user increasing from time to time however the interested people are new and inexperienced on Weka usages therefore they faced hard to identify the best approach for their particular dataset among the many available Weka Algorithms [123]. So this new version of Auto-WEKA; a system designed to help such users by automatically searching through the joint space of WEKA's learning algorithms and their respective settings to maximize performance, it uses Bayesian optimization method by integrating with this new package in Weka to make it just as accessible to end users as any other learning algorithm. The below figures shown that; the "Random Forest " is significantly performing higher than any other algorithms to be applied on model selection of EDHS 2016 datasets.

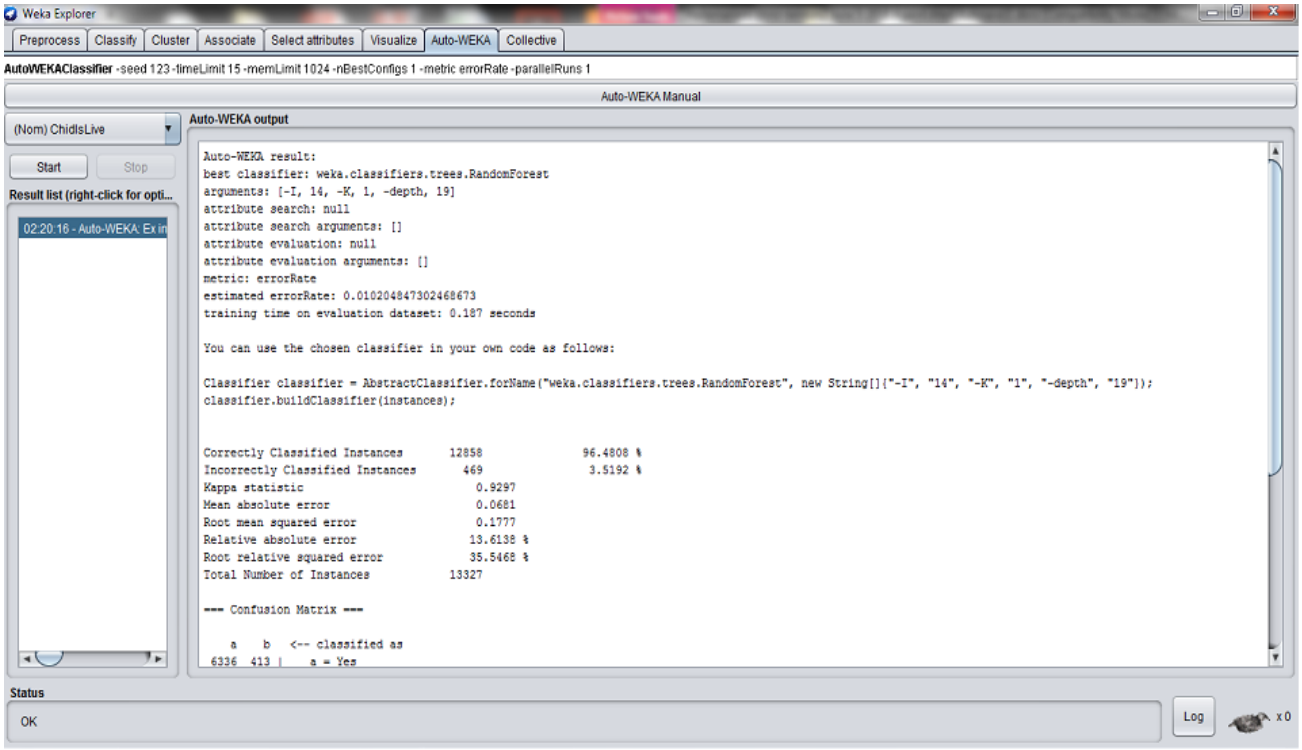


Figure 4. 2: Auto-Weka classifiers Selection Results (A)

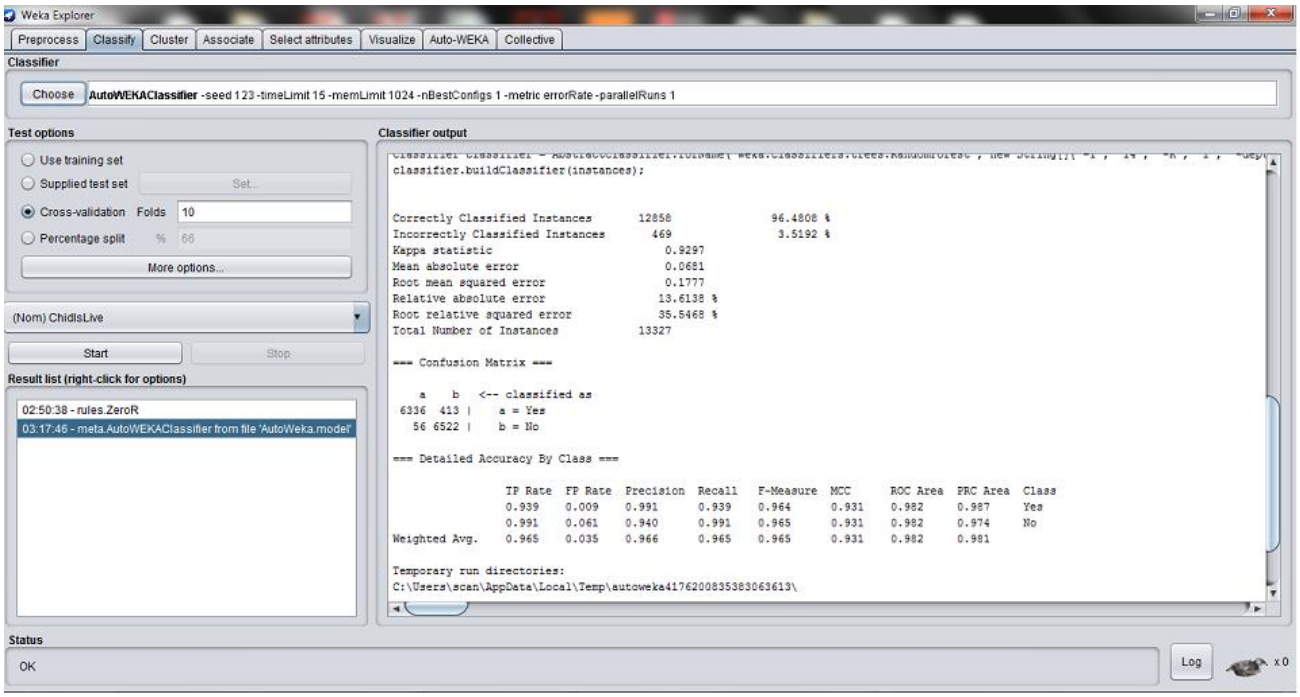


Figure 4. 3: Auto-Weka classifiers Selection Results (Confusion Matrix) (B)

4.5.2. Choosing classifiers using Cost/Benefits Analysis

4.5.2.1. Building the PART Decision Lists

Build a PART and Random Forest model for predicting the child mortality among the total children ever born; so this is to identify statistically significant classification model for predicting mortality as well as to select the more suitable classification model for the datasets of EDHS 2016.

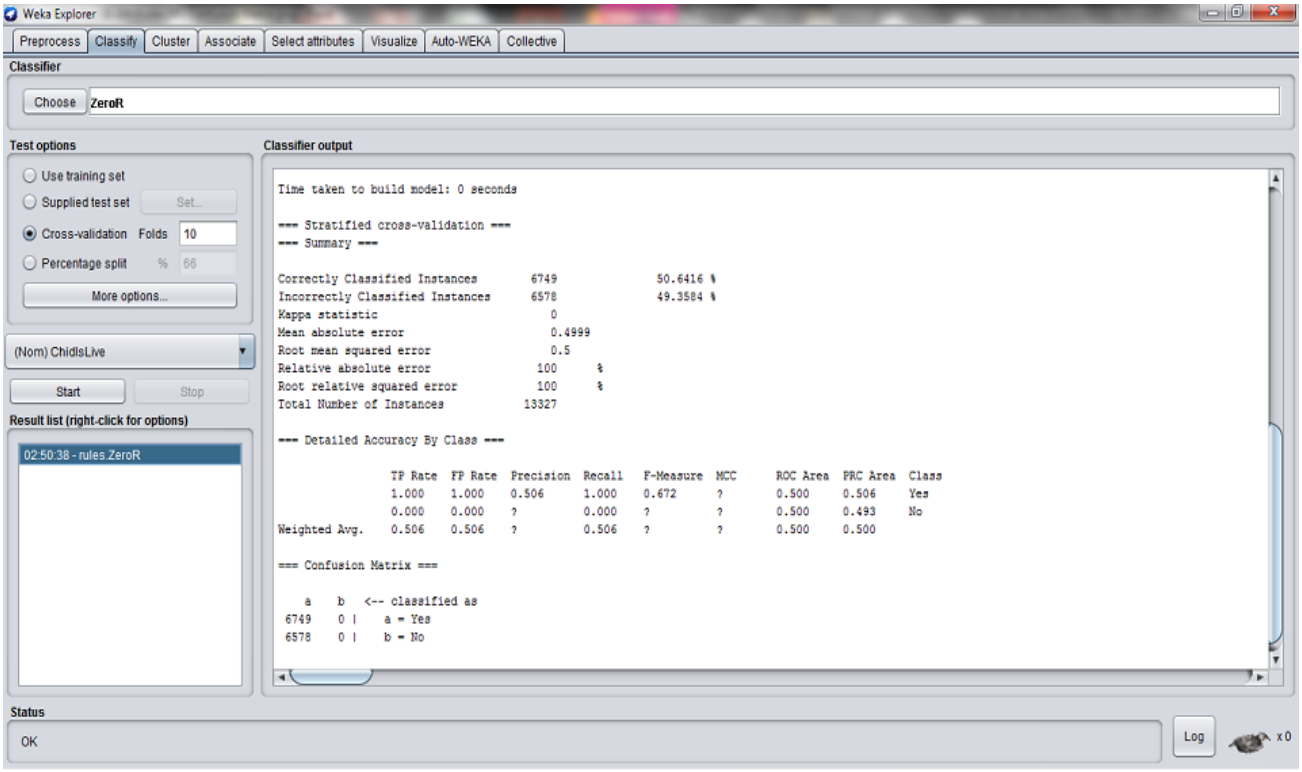


Figure 4. 4: The performance of ZeroR

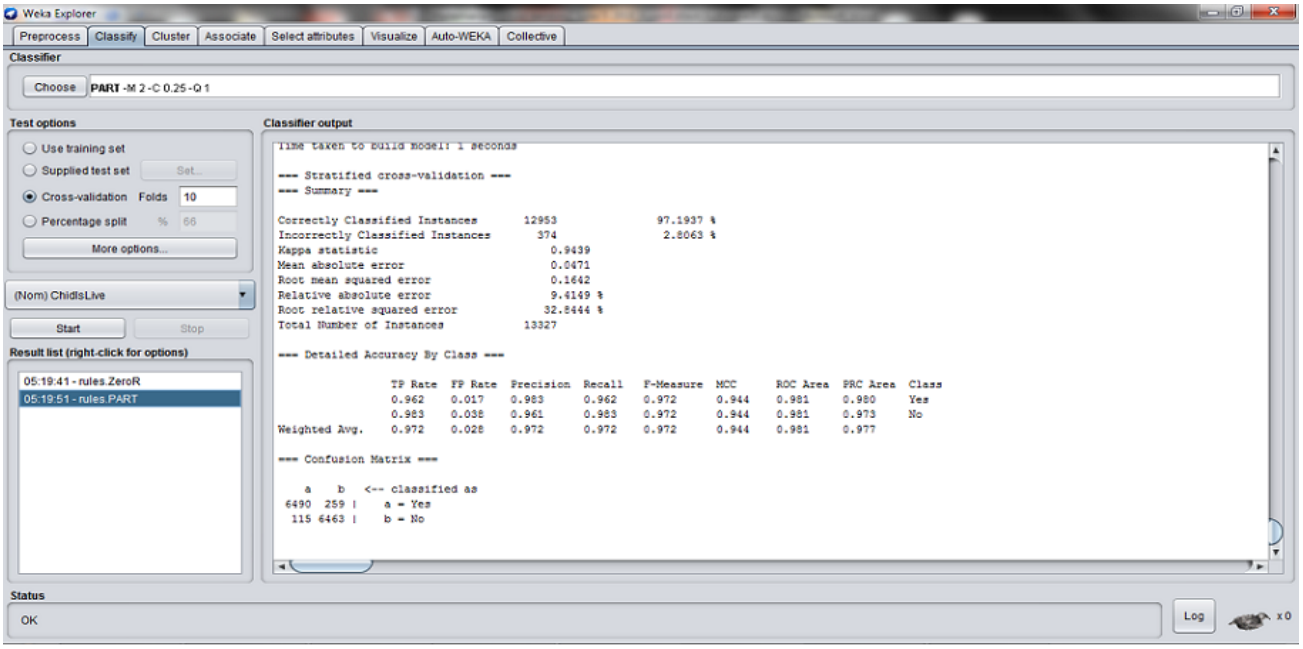


Figure 4. 5: The performance of PART algorithms

The accuracy of the model became higher (97.1937% instead of 50.6416%), its real statistical significance became much stronger. This follows from the value of the ''Kappa statistic" 0.9439, which indicates the existence of moderate statistical dependence. It can be analyze using the "Confusion Matrix" at the bottom of the Classifier output window. Therefore, there are 6,463 true positive, 6,490 true negative, 259 false positive, and 115 false negative, and 259 false positive children classified wrongly. It is because of the considerable number of false positive that the value of precision for "died-child (No)" mortality 0.961 is rather higher than ZeroR. Nonetheless, the model exhibits an excellent value of "ROC Area" for "died-child (No)" cases 0.9808. This indicates that PART model could be use very advantageously for discovering (died-child (No)) through virtual screening. This can clearly show by analyzing ROC and Cost/Benefit plots. The PART method provides estimated probability in classification tasks. This means that PART models can assess the value of the probability (varying from 0 to 1) that a given cases can be predicted as "died-child (NO)". By moving the threshold from 0 to 1 and imposing that a case can be predicted as "died-child (No)" if the corresponding probability exceeds the current threshold, one can build the ROC (Receiver Operating Characteristic) curve.

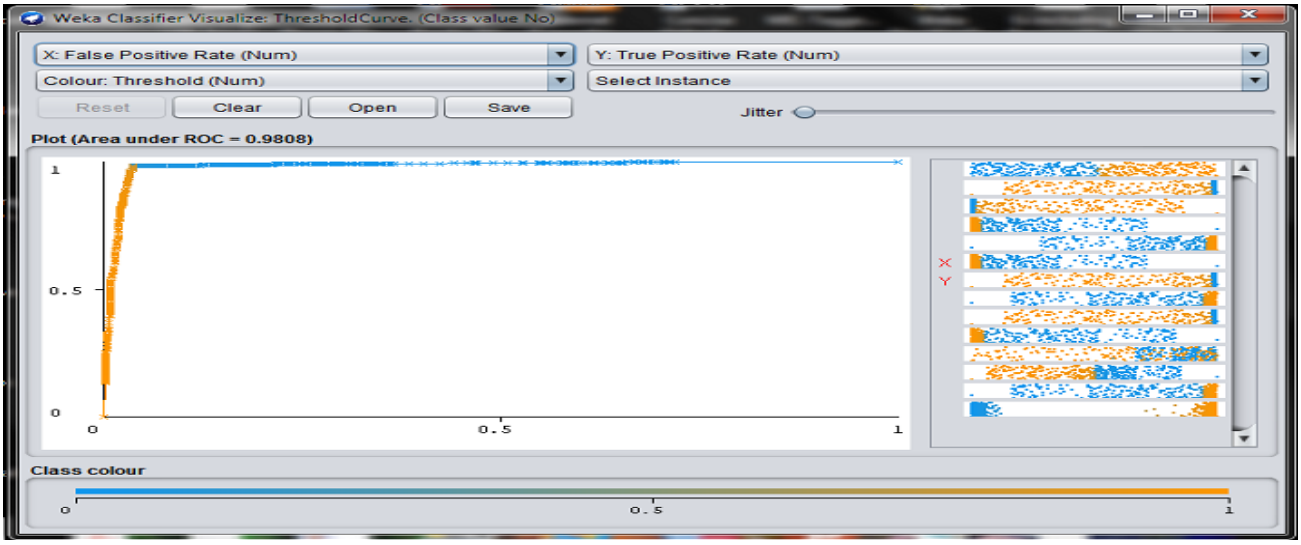


Figure 4. 6: Threshold curve of ROC area of child died (No)

The axis X in it corresponds to the false positive rate, whereas its axis Y corresponds to the true positive rate. The color depicts the value of the threshold. The "colder" (closer to the blue) color corresponds to the lower threshold value.

The color showed the value of the threshold. The color closer to the blue color related to the lower threshold value (look class color ranges). All cases with probability of being "died-child (No)" exceeding the current threshold is predicted as "died-child". If such prediction made for a current case is correct, then the corresponding case is true positive, otherwise it is false positive. If for some values of the threshold the true positive rate greatly exceeds the false positive rate, then the classification model with such threshold can be used to extract selectively "died-child (No)" cases from its mixture with the big number of "Yes (child is Live)" ones. In order to find the optimal value of the threshold can perform the cost/benefit analysis.

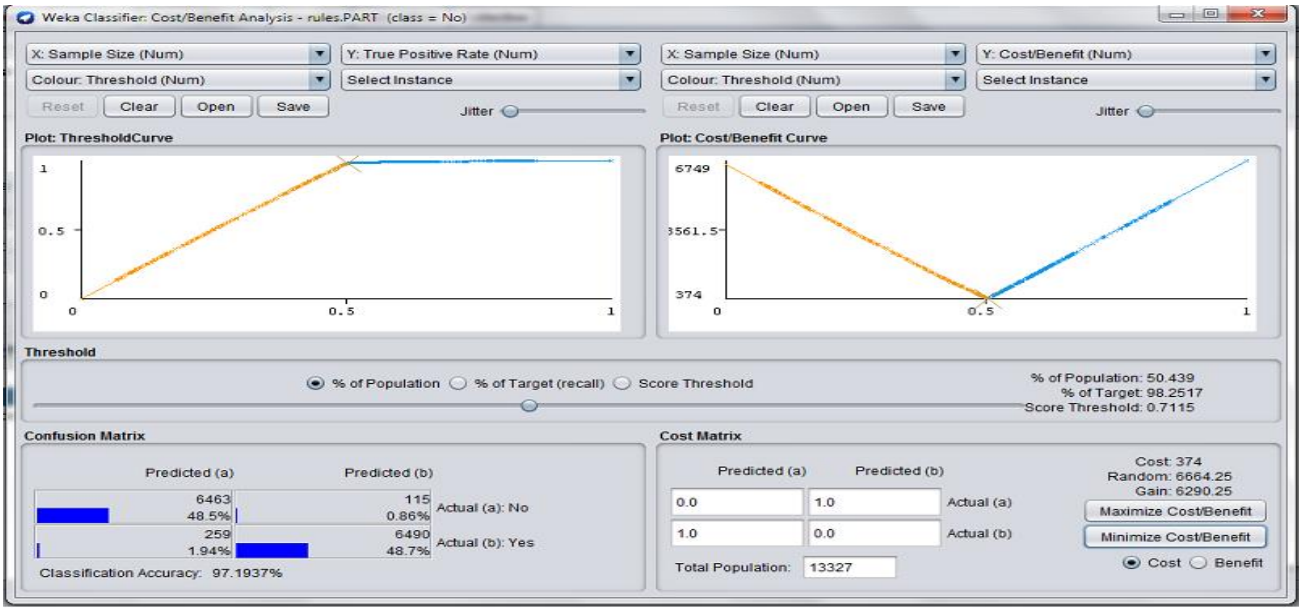


Figure 4. 7: The cost /Benefits analysis on PART algorithms

The Threshold curve looks very similar to the ROC curve. In both of them, the axis Y corresponds to the true positive rate. However, in contrast to the ROC curve, the axis X

in the threshold curve corresponds to the part of selected instances (the "Sample Size"). In other words, the Threshold curve depicts the dependence of the part of "Died-child (No)" cases retrieved upon the part of cases selected from the whole dataset used for identifying. The confusion matrix for the current value of the threshold shown in the Confusion Matrix frame at the left bottom corner of the window.

Its current value is 374 (cost of 115 false positives and 259 false negatives). In order to find the threshold corresponding to the minimum cost, it is sufficient to press the button Minimize Cost/Benefit. The initial confusion matrix corresponds to the threshold 0.5, whereas the second confusion matrix results from the value of the threshold found by minimizing the cost function. The difference between the cost values and random selection called "Gain". The Gain can be interpret as the profit by using the classification model instead of random selection of the same number of child mortality cases.

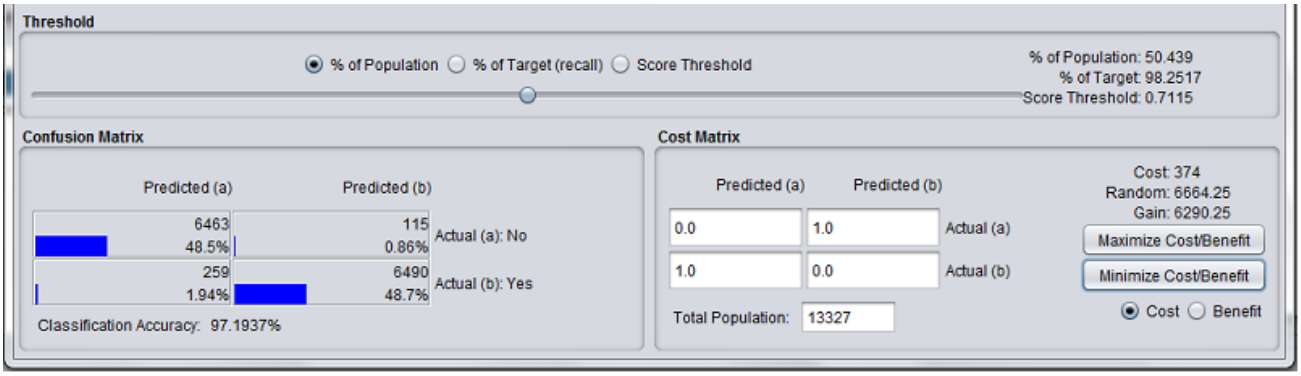


Figure 4. 8: Minimize Cost /Benefits analysis on PART algorithms

This is the sample ROC cure conducted for PART algorithms by Cost /benefits analysis.

4.5.2.2. Building the Random Forest Model

Building the models using PART decision list are not very strong from statistical results so that it can be strongly improve by applying ensemble (Random Forest) modeling. In the latter case, an ensemble of several models built as an alternative of a single one, and

prediction of the ensemble model made as a consensus of predictions made by all its individual members. Random Forest is the dominant algorithms that used on the ensemble modeling for this datasets.

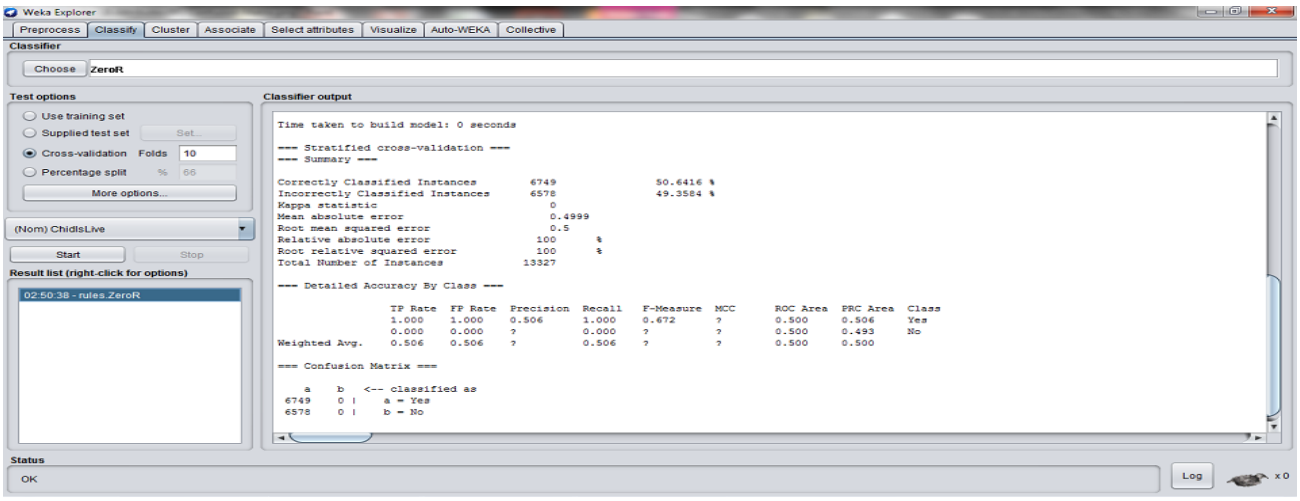


Figure 4. 9: The performance of ZeroR

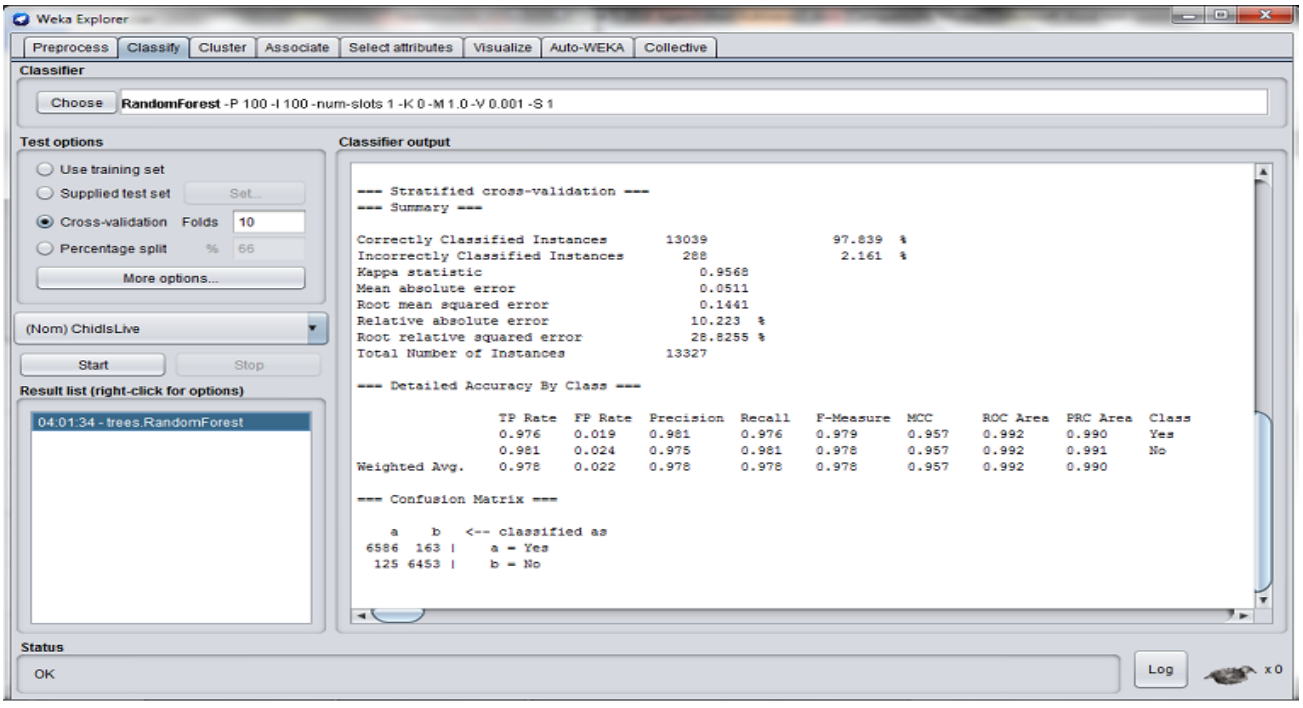


Figure 4. 10: The performance of Random Forest algorithms

Random Forest model could be use very advantageously for discovering died-child (No) through virtual screening. This can clearly show by analyzing ROC and Cost/Benefit plots. The Random Forest method provides estimated probability in classification tasks. This means that Random Forest models can assess the value of the probability (varying from 0 to 1) that a given cases can be predicted as "died-child (NO)". By moving the threshold from 0 to 1 and imposing that a case can be predicted as "died-child (No)" if the corresponding probability exceeds the current threshold, one can build the ROC (Receiver Operating Characteristic) curve.

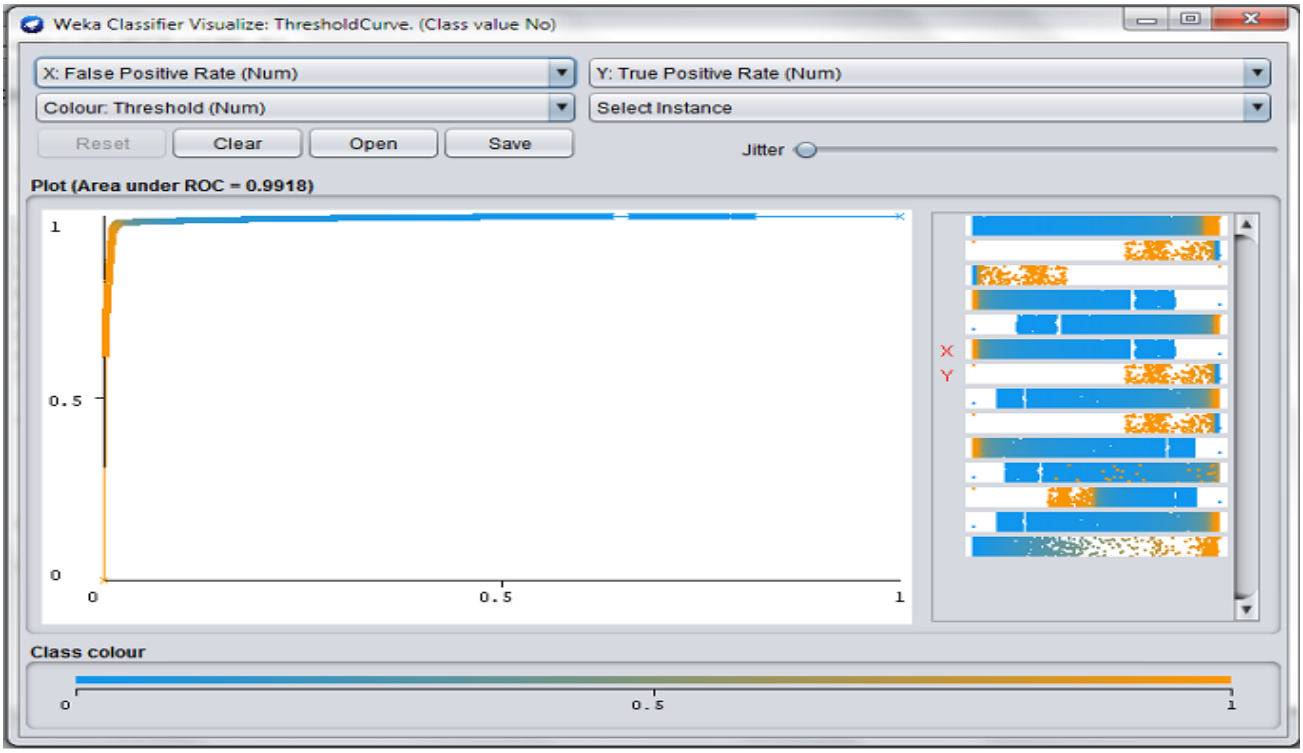


Figure 4. 11: Threshold curve of ROC area of child died (No)

The color showed the value of the threshold. The color closer to the blue color related to the lower threshold value (look class color ranges). All cases with probability of being "died-child (No)" exceeding the current threshold is predicted as "died-child". If such prediction made for a current case is correct, then the corresponding case is true positive,

otherwise it is false positive. If for some values of the threshold the true positive rate greatly exceeds the false positive rate, then the classification model with such threshold can be used to extract selectively "died-child (No)" cases from its mixture with the big number of "Yes (child is Live)" ones. In order to find the optimal value of the threshold can perform the cost/benefit analysis.

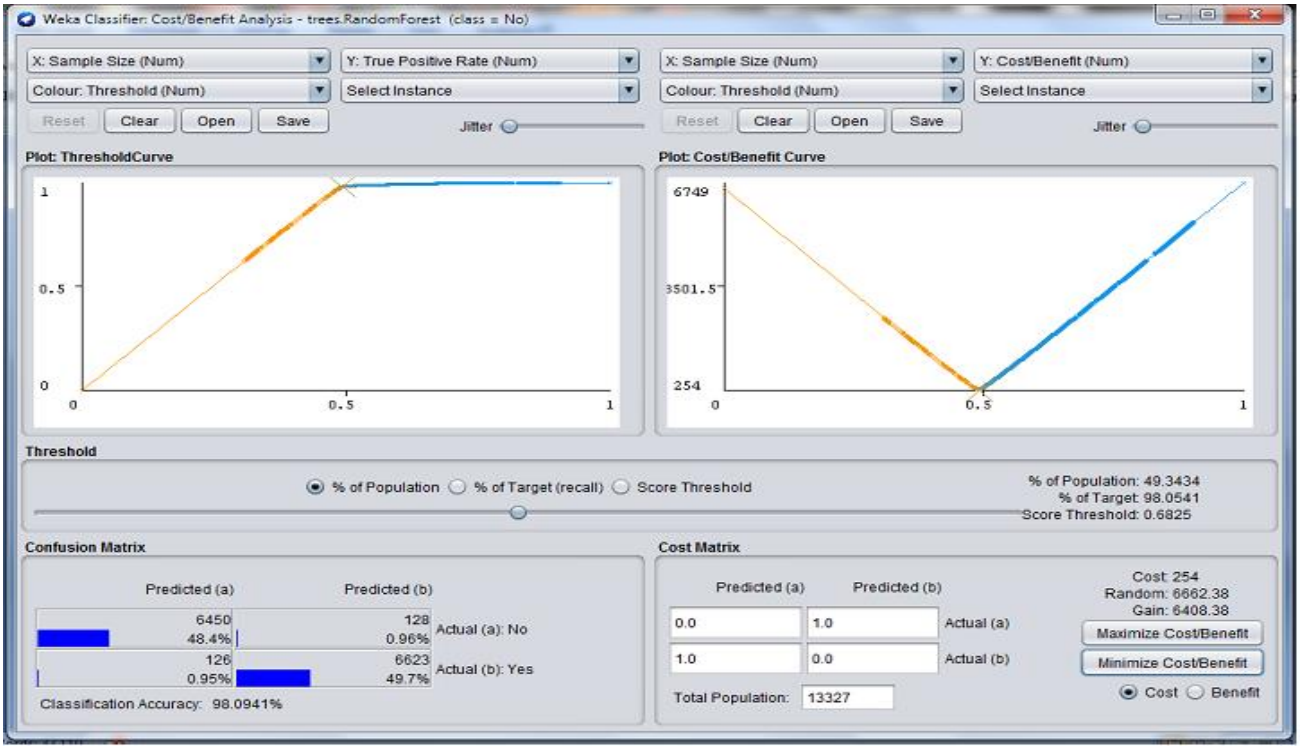


Figure 4. 12: The cost /Benefits analysis on Random Forest algorithms

The confusion matrix for the current value of the threshold shown in the Confusion Matrix frame at the left bottom corner of the window. The confusion matrix for the current value of the threshold sharply differs from the previously obtained one. In particular, the classification accuracy 98.0941% is considerably higher than the previous value 97.839%, the number of false positives has greatly decreased from 163 to 128, and the number of false negatives has slightly increased from 125 to 126. Its default value is 1 unit. In the case of virtual screening, this corresponds to the mean price one should pay for "neglects"

very useful case and losing profit because of the wrong prediction taken by the classification model. It also taken by default that one should not pay price for correct decision taken using the classification model.

Its current value is 254 (cost of 128 false positives and 125 false negatives). In order to find the threshold corresponding to the minimum cost, it is sufficient to press the button Minimize Cost/Benefit. The initial confusion matrix corresponds to the threshold 0.5, whereas the second confusion matrix results from the value of the threshold found by minimizing the cost function. The difference between the cost values and random selection called "Gain". The Gain can be interpret as the profit by using the classification model instead of random selection of the same number of child mortality cases.

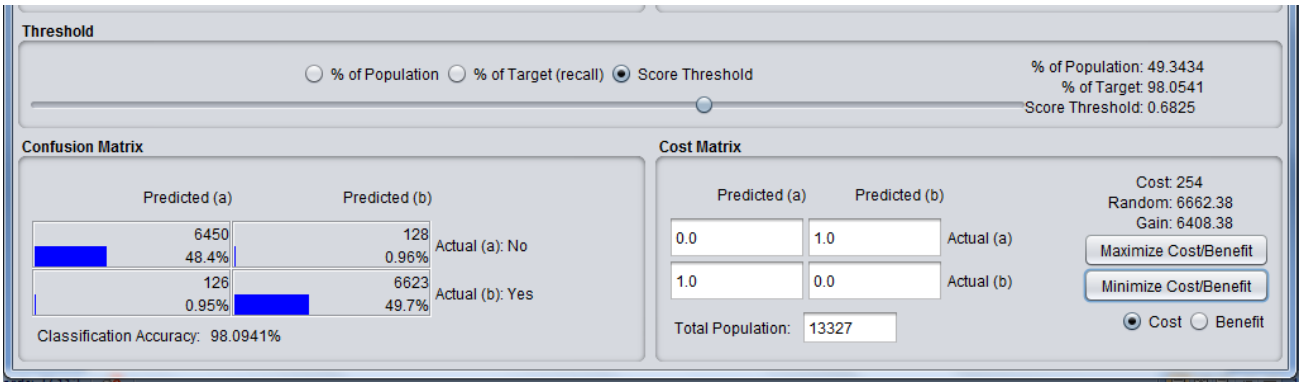


Figure 4. 13: Minimize Cost /Benefits analysis on Random Forest algorithms

This is the sample ROC cure conducted for Random Forest algorithms by Cost /benefits analysis.

The accuracy of the Random Forest classification and kappa statistics value are higher compare with PART; however, the ROC area higher coverage so this indicate that it is a sign of a good opportunity virtual screening and we can used it to analyze the cost/benefit of the model.

The cost/benefit analysis show that it has a very good parameter observation that means the cost decreasing, the Gain is increased and to select the "died-child (No)" children 29.8575% of cases are enough to retrieve 95.6044% within 0.7506 score threshold.

Finally, we can conclude that from the result of Section 4.5.2.1 and Section 4.5.2.2. Point of view "random forest" algorithms selected for the datasets of EDHS 2016 as the best suitable classifiers of the child morality cases.

Chapter Five

Experimentation and Analysis

5.1. Overview

In this experimentation and analysis, phases will be discuss about the models building and experimentations conducted. Model building is an iterative process so hybrid KDP methodology was select for this study. Moreover, to achieve the main objective of this study the number of experiments conducted based on the datasets. Therefore, it is important to conduct different experiments to find the best model for solving the problems.

The main objective of the study was to predict the pattern of under-five mortality in Ethiopian. Therefore, to achieve this data mining classification technique applied to develop predictive models.

There are no best classifiers for all problems in this world but you may find different classifiers that generated similar result[124]. For this study, several classifiers compared in this study and the result of the classifiers performance examined during cross-validation phases.

Although the PART classifiers had reasonable accuracies in the validation phases, the RF delivered the most stable performances (99%), regardless of applied or absent feature preprocessing. Other classifiers had varying, lower results.
In general there are number of experiments were conducted using the algorithms to predict whether child is Lives or death. Therefore, Random Forest selected and experimented on EDHS 2016 datasets using Weka 3.8.2 version software. Moreover, different experiments conducted in each of the outcome variables and evaluated their performance with the output from the algorithms.

5.2. Experimental Design

Classifiers have different performances when used in different datasets. So experiments tested for each classifier using the EDHS 2016 dataset consisting of 6,893 instances and 18 attributes including the class attribute (dependent variables). A stratified 10-fold cross-validation used to estimate the performance of random forest classifier on each of parameters. In the above on chapter 2 the performance estimation approach has been proved to be statistically good enough in evaluating the performance of data mining classifier algorithms. Overall, now classification Accuracy, TP rate, FP rate, Size of the tree, and ROC area used to evaluate and compare the performance of the models generated. These measures driven from the confusion matrix of the models. The algorithms used for predictive model building found in Weka 3.8.2. This version works on many file formats and it is compatible with CSV file format. This can be change easily from access (csv) to .arff file format, which is available to save in Weka format.

5.3. Model Building Using the Random Forest

To build the predictive model the selected datasets were given to Weka in .ARFF and/or CSV files. In the above, as mentioned for this study Random forests predicting algorithms are used.

According to chapman definition " 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"[52]. Therefore, in this study will generate classification rules that were assigning the correct class label to previously unseen and unlabeled children.

Classification is mapping (classifies) a data item in to one of several predefined classes. Due to this one of the classifiers selected is Random forest because the well-developed

randomizing more useful than any other mapping algorithms[99]. And the other main advantages are the model produced by random forest easily can be interpreted based on the visualization effects and manually can build using if conditions[56].

For most of the experiments carried out in this phase, the experiments were 10-fold cross validation were used, the total record was partitioned in to 10-folds, the training and test datasets. These two datasets found from the final dataset by using a stratified sampling technique where the different classes, found in the classification, were considered as strata for 10-fold cross validation

Different experiments conducted for the Random forest classifier by changing the main parameters of the algorithm to build a better predictive model. As displayed in the tables, different experiments carried out by using all the 18 attributes of the records with different schemes applied in the experiment and two different test modes (ways of feeding records to the algorithms).

Analysis of the Random forest 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 experiments for Random Forest classification models listed under beneath Table 5.1, Table 5.2.

Experiment 1 Number of Iteration is 10 with 10-fold cross Validation
Experiment 2: Number of Iteration is 50 with 10-fold cross Validation
Experiment 3: Number of Iteration is 100 with 10-fold cross Validation
Experiment 4: Number of Iteration is 150 with 10-fold cross Validation
Experiment 5: Number of Iteration is 50 with 66%, split tests mode
Experiment 6: Number of Iteration is 50 with 70%, split tests mode
Experiment 7: Number of Iteration is 50 with 80%, split tests mode
Experiment 8: Number of Iteration is 50 with 90%, split tests mode

S. No	Comparing parameters	Experiment Number			
		Experiment1	Experiment2	Experiment3	Experiment4
1	Number of iteration	10	50	100	150
2	Classification Accuracy %	97.7714	97.854	97.839	97.8465
3	Precision	0.974	0.976	0.975	0.975
4	Recall	0.981	0.981	0.981	0.981
5	F-Measure	0.978	0.978	0.978	0.981
6	ROC area Curve	0.990	0.992	0.992	0.992

Table 5. 1: The result of Random Forest with 10-fold CV test mode and the Number of Iteration.

As we observed from Table 5.1, the number of iteration changed then the performance of the classification accuracy, precision, Recall, F-measure and ROC also changed. so as an investigator can be inferred from the above Table 5.1 that experiment 2 has higher accuracy, precision and other performance measurement has similarity than experiments 1,3&4 because the performance of correctly classifying of instance decreased from 97.854%(50 Number of Iteration) to 97.839% and 97.8465% (100&150 Number of Iteration). Therefore, experiment 2 outperforms than the other experiments in performance for EDHS 2016 datasets.

The data collected, preprocessed and analyzed using classification (Random Forest) 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 %
Experiment 5	66%	97.8813
Experiment 6	70%	97.924
Experiment 7	80%	97.9362
Experiment 8	90%	97.4494

Table 5. 2: The Results of Random Forest with different percentage split test mode.

As it can be observe from the above table, the 80% split test of data for training is better than the other percentages split test options. The selected percentage has 97.9362% correctly classified instances. Percentage split test parameter of 66% and 70% training set

also have 97.8813% and 97.924% correctly classified instances respectively, but with relatively low precision, recall, F-measure and ROC curve.

Moreover, the Random Forest model produced from the Table 5.2 experiment 7, which has 80% 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 is 97.9362%. The 80% test option mode shows a better performance than others do.

The percentage split test option used to partition the dataset into training and testing data and this parameter was set to 80, i.e. 80% for training and 20% for testing. The result of this learning scheme summarized and presented below in Table 5.3 with respective performance matrices values, accuracies, number of iteration, size of trees, and ROC/AUC curve.

Let test the Inputting all the records Percentage(80% training and 20% test) split test which is train a model and then supply the unseen remaining part of the record for testing the performance of the model using the following experiments.

Experiment 9: Number of Iteration is 10 with (80% training and 20% test), split tests mode
Experiment 10: Number of Iteration is 50 with (80% training and 20% test), split tests mode
Experiment 11: Number of Iteration 100 with (80% training and 20% test), split tests mode

S.No	Comparing parameters	Experiment Number		
		Experiment 9	Experiment10	Experiment11
1	Number of iteration	10	50	100
2	Classification Accuracy %	99.1745	99.3246	99.3246
3	Precision	0.987	0.988	0.988
4	Recall	0.996	0.998	0.998
5	F-Measure	0.992	0.993	0.993
6	ROC area Curve	0.999	0.999	0.999

Table 5. 3: The resulting Random Forest with (80%training and 20% tests) test mode

As can be observed from this Table 5.3, experiments 10 have comparatively better than other experiments to extracted accuracies and rules. This is because, the researcher used to build Random Forest with 80% training and 20% testing splits with different number

of iteration factors and also to choose the best from all experiments are the classification accuracy are more considerable than other factors. This result described in particular ways due to the adjustment of some of the parameters,

The model has accuracy of 97.854% using 10-fold cross-validation and 99.3246% accuracy using 80% split test options. Moreover, the model has a true positive rate of 98.1% and true negative rate of 97.6% for 10-fold cross-validation and a true positive rate of 99.8% and true negative rate of 98.8% for 80% split test options.

The best Random Forest model of the classification generated from experiment 10 of the 80% split test mode. The model shows a better performance evaluation than other models. The 80% 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 10 with 80% split test mode options, which is Random Forest with 50 number iterations, are statistically significant in splitting the decision tree. Furthermore, reviewing of literatures indicated that these classifiers (Random Forest) have a great role in the prediction tasks.

Therefore, to conclude this below will mentioned some of reasons why experiment 10 chooses from other experiment as follows;

- ⇨ Relatively the numbers of records are large.
- ⇨ The attribute selected effectively used.
- ⇨ The size of the tree on experiment 10 are manageable; and
- ⇨ The number of rules extracted in experiment 10 is reasonable.
- ⇨ 80% of split tests mode are used in experiment 10 is acceptable.
- ⇨ Relatively the accuracy of the resulting model is best than others.

5.3.1. Confusion Matrix for Random Forest Model

To understand easily about the accuracy and errors of the model confusion matrix more suitable; in addition, it provides the overall accuracy of the prediction of unseen instances and it helps to observe the explanatory analysis of the performance of classifiers.

Below will presents the confusion matrix for Random forest Table19 showed that out of the total records provided to the program 1330 (98.81%) and 1317(99.77%) records were classified correctly in the class of Alive and Dead respectively. On the other hand, 16 (0.23%) records were incorrectly classified as "Dead" while actually, they were supposed to be in the "Alive" class and 2 (1.19%) records were classified incorrectly as Alive while actually they are in the Dead class. This explained that from the total records 2665 (99.32%) records classified correctly while the remaining 18 (0.68%) records classified incorrectly. Hence, this indicated that records whose class is "Alive" classified with a minimum error as compared with the records in the class "Dead".

Actual Class	Predicting Class			Total
		Alive	Dead	
	A live	1330	16	1346
	Dead	2	1317	1319
	Total	1332	1333	2665

Table 5. 4: Confusion Matrix for Random Forest model

5.3.2. ROC Analysis for Random Forest Model

A ROC curve is the means of comparison between individual models and shows easily the high proportion of positive target. In the below Figure 5.1 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.

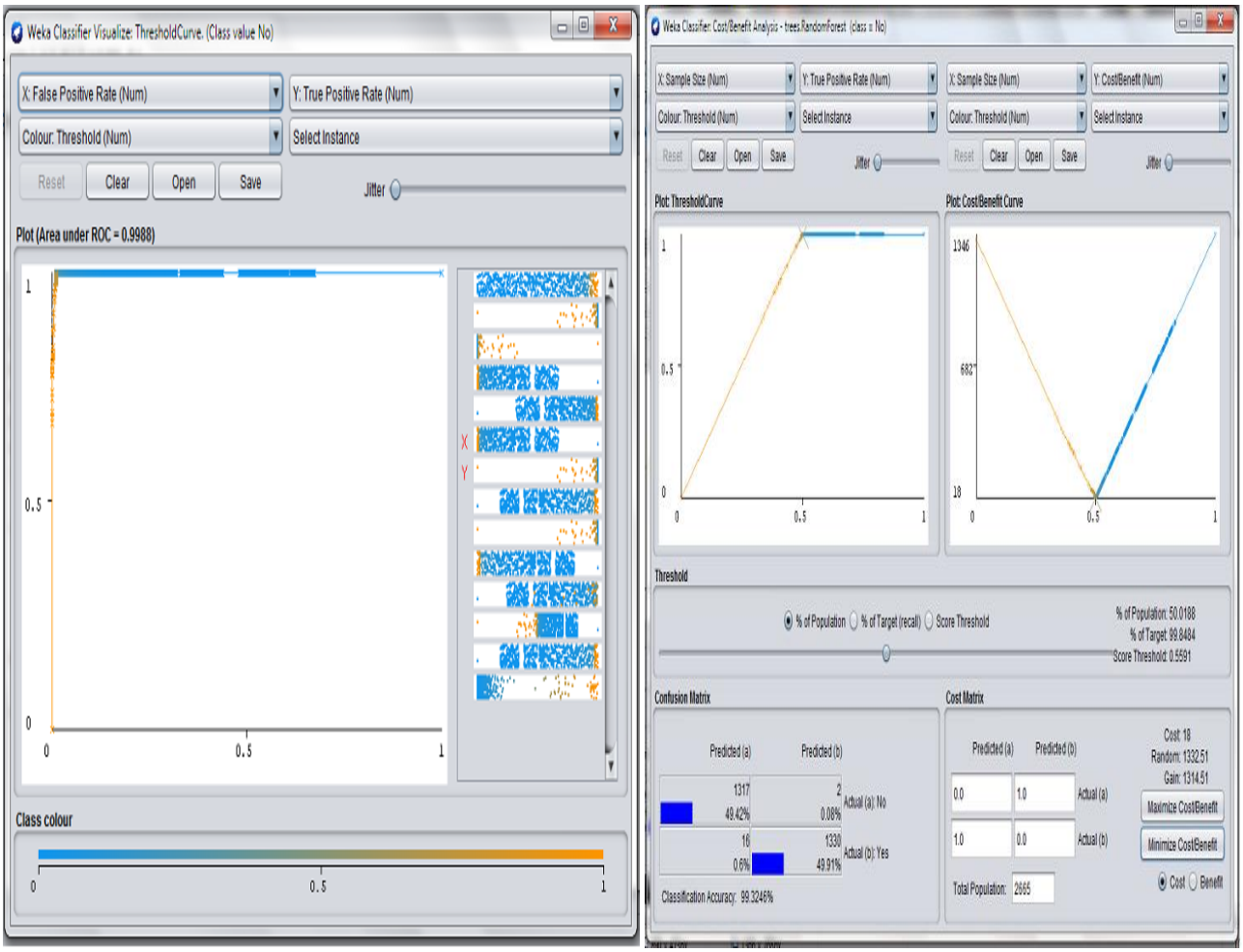


Figure 5. 1: ROC curve of the Random Forest model with 20% Test set

The true positive (TP) rate and false positive (FP) rate values of different classifiers on the same test set often represented diagrammatically by a ROC Graph. The abbreviation ROC analysis stands for 'Receiver Operating Characteristics Graph', ROC analysis related in a direct and natural way to cost/benefit analysis of decision-making. Class value Alive gives the ROC accuracy of 98.88%. The larger the area under the ROC curve (AUC), the higher the likelihood that an actual positive case and the better the model. The AUC for the model is 0.9898, which is closer to one (1) that is AUC measure is especially useful for datasets with unbalanced target distribution (where one target class dominates the other).

5.4. PART Classifier Model Building using WEKA Software

The second type of classification technique applied in this study is the PART Rule algorithm.

PART is a separate-and-conquer rule learner proposed by Eibe and Witten[15]. PART is an indirect method for rule generation that means the algorithm producing sets of rules called decision lists, which ordered set of rules. A new data is compare to each rule in the list in turn, and the item assigned the category of the first matching rule (a default applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning[94, 95]

PART is Weka's implementation of the C4.5 algorithms, which can work on multiple valued attributes. It contains some parameters that can be change to improve its performance on different values. Different experiments conducted for the PART classifier by changing the main parameters of the algorithm to build a better predictive model.

Parameter	Description	Types
BinarySplits	Whether to use binary splits on nominal attributes when building trees	Boolean
Confidence Factor	The confidence factor used for pruning (smaller values incur more pruning)	Numeric
MinNumObj	The minimum number of instances per leaf	Numeric
unpruned	Whether pruning is performed	Boolean

Table 5. 5: The Parameters of the selected PART Rule classifier

BinarySplits parameter by default is set to False. If this value is changed to True, it enforces the model generated to be binary decision tree rather than generalized decision tree. The confidence factor helps to set a limit so that the algorithm makes more or less pruning. The default value for confidence factor is 0.25. The minimum number of instances per

leaf is set 2 and the last parameters whether pruning is performed or not is defaulted to False mainly affect the performance of the model built. Moreover, it is to see the effect of tree pruning method on classification accuracy of PART, the Rule size, and the time taken to build the model by the algorithm. In addition, the 10-fold (cross validation test) option conducted. Moreover, it is to see the effect of tree pruning method on classification accuracy of PART Rule, the tree size, and number of leaves to build the model by the algorithm. In addition, the 10-fold (cross validation test) option employed.

So using PART Rule eight (8) experiments conducted using all attributes for each of the two dependent variables. In the experiment, the 10-fold cross validation used for all experiments. The PART Rule experiment designed to build the model for predicting child mortality pattern and to compare the performance with Random Forest algorithm.

The following experiments was designed to build a model by changing the values of parameter as shown in Table 5.6 for each of the outcome variables used for this study and this reduced table are produced after 6X4 tables performance are reviewed here.

	Parameters		
Experiments	**Un-pruned**	**Confidence Factor**	**Numbers of Instance(minNumObj)**
Experiments1	True	0.15	2
Experiments2	True	0.25	2
Experiments3	True	0.30	2
Experiments4	True	0.50	2
Experiments5	False	0.15	2
Experiments6	False	0.25	2
Experiments7	False	0.30	2
Experiments8	False	0.50	2

Table 5. 6: Values of parameters used in these experiments

After all necessary PART Rule algorithm parameters were set as shown on above Table 5.6. Then the eight experiment listed in Table 5.7 will be conducted by the sets of

parameters and the summary of the result of the eight (8) experiments is presented in Table20.

	Experiments							
Performance measure	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
Number of Rules	395	395	395	395	96	105	104	**105**
Classification Accuracy %	96.819%	96.819%	95.819%	96.819%	97.216%	97.194%	97.261%	**97.321%**
Precision	0.954	0.954	0.954	0.954	0.962	0.961	0.963	0.964
F-Measure	0.968	0.968	0.968	0.968	0.972	0.972	0.973	0.973
ROC area Curve	0.972	0.972	0.972	0.972	0.981	0.981	0.981	**0.981**

Table 5. 7: Summary of PART classifier Experiment Results using 80% of Split tests

As can be observed from the above Table 5.7, the model scored in experiment (1-4) with un-pruned all parameter has constant value consecutively which tested on 80% of split test. This means that, PART rule algorithms are more convenient when using pruned tree than un-pruned ones. However, when reviewing the overall performance of the models, the maximum performance result was achieved when the value of the confidence factor and number of Rules created at (0.50, 105) and (0.30, 104) respectively so that experiment 8 chosen because of the more number of Rule are the better than lesser. The second choosing criteria is higher accuracy value which is 97.321%[87]

5.4.1. Confusion Matrix for PART Rules Classifiers

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

Confusion matrix is a model classification conducted based on the test result. The overall purpose of confusion matrix is predicting the accuracy of unseen instances it is often helpful to see a collapse of the classifier's performance. Below Table20 shows the PART

rules form of a confusion matrix that 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 and 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	Predicting Class			Total
		Alive	Dead	
	A live	1315	31	1346
	Dead	22	1297	1319
	Total	1337	1328	2665

Table 5. 8: Confusion Matrix for PART Rule model

The confusion matrix for PART classifier presented in the above table20 depicts that out of the total records provided to the program 1,315 (97.69%) and 1297(98.33%) records were classified correctly in the class of Alive and Dead respectively. On the other hand, 31(2.31%) records were incorrectly classified as 'Dead' while actually they were supposed to be in the 'Alive' class and 22 (1.67%) records were classified incorrectly as 'Alive' while actually they are in the 'Dead' class. This displayed that from the total records 2,612 (98.01%) records classified correctly while the remaining 53(1.99%) records classified incorrectly. Hence, this indicated that records whose class is 'Alive' classified with a minimum error as compared with the records in the class 'Dead'.

5.4.2. ROC Analysis for PART Rule Classifiers

ROC area curve is generated by drawing curves in two-dimensional spaces, with X and Y-axes defined by the True Positive rate and False Positive rate. The AUC for the under-five mortality records generated from the PART Rule Classifier is presented in the below Figure 5.2 ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. Class value dead gives the ROC accuracy of 99.09%.

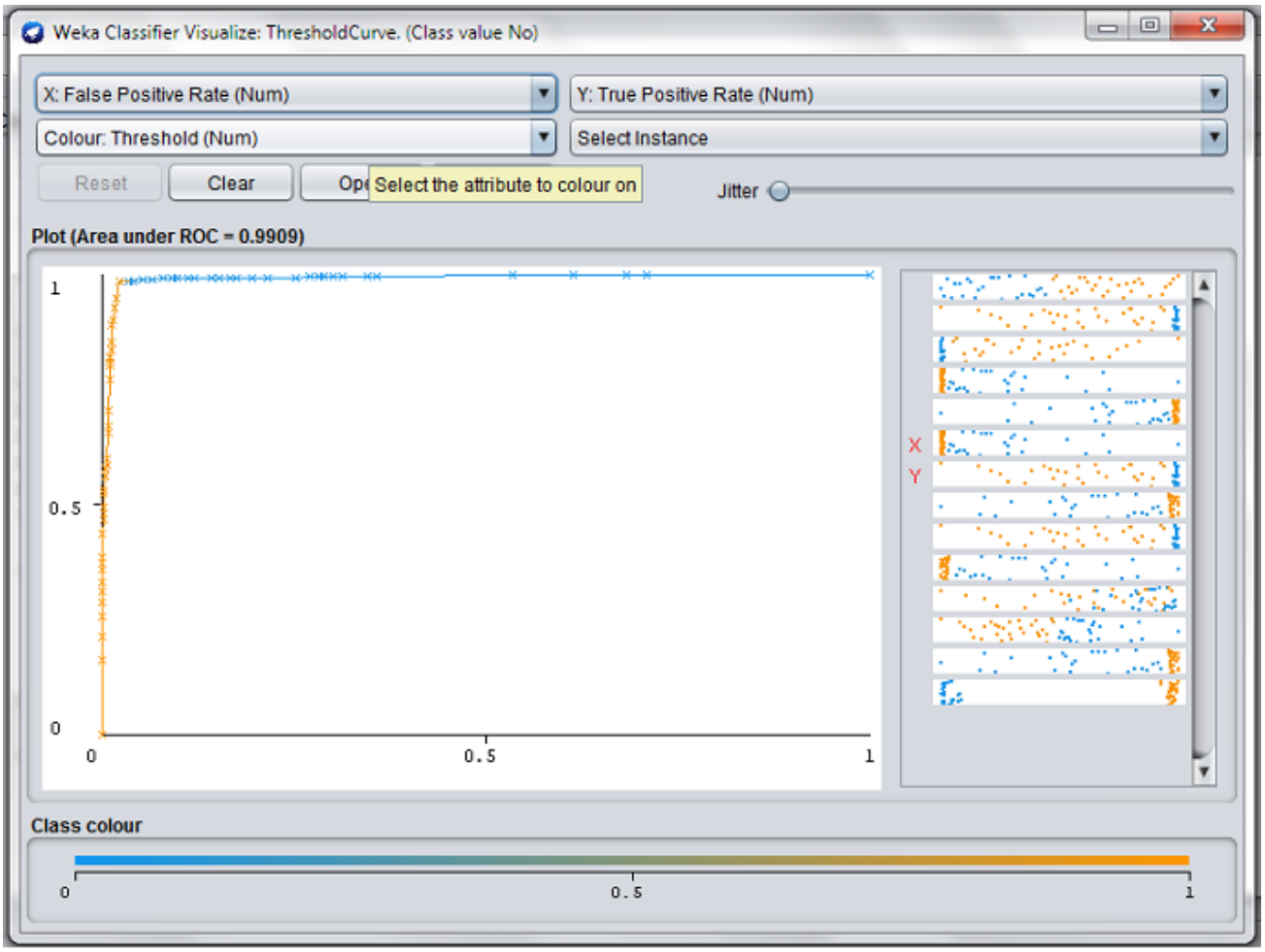


Figure 5. 2: ROC curve from the PART Rule Classifier

In the above Figure 5.2, 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).

In Figure 5.2, the area under the diagonal curve is 0.9909. 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.

5.5. Performance comparison of the Classification Model

In this research work, several experiments had been carryout with two classification algorithms, i.e. Random Forest and PART Rule algorithm to build a predictive model that predicts the pattern of under-five mortality in Ethiopia, particularly for EDHS 2016 datasets. From the experiments, all attributes identified to make sound rule and better accuracy. Both classifiers algorithms compared due to inputting 80% 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 Random Forest algorithm and PART Rule classifiers are summarized in Table21 the predicting of the under-five mortality, accuracy, precision (for two class dataset), Area under the ROC, have been computed to give a deeper insight of the automatic diagnosis. Moreover, a comparison of the performance evaluation in the Table 5.9 between Random Forest and PART Rule decision list algorithm illustrated in below.

Performance Testing	Random Forest	PART Rule
Accuracy (%)	99.32%	98.011%
Precision (%)	0.988	0.977
Recall (%)	0.998	0.983
F-Measure (%)	0.993	0.980
AUC (%)	99.99%	99.09%

Table 5. 9: Performance comparison of Random Forest and PART Rule classifier with 80% split test mode

The purpose of the above Table 5.9 constructed to select the best performing classifiers from the two algorithms (Random Forest or PART Rule) used for data mining models. Therefore, according to the comparison result of the two classifiers after applied for EDHS

2016 datasets. Based on this, the Random Forest has been select than PART Rule algorithms because this algorithm has higher accuracy than PART Rule.

In addition to this as you, observe that the time taken to execute the experiment PART is faster than Random Forest. However, the overall result scores of the Random Forest algorithm is higher than that of the PART Rule classifier model. In this study, the models evaluated based on the accuracy measures discussed above (classification accuracy, Precision, Recall, F-measure and AUC). The results achieved using inputting 80% 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 0.9932 with Precision of 0.988, Recall is 0.998, AUC is 0.999, and F-measure are 0.993.

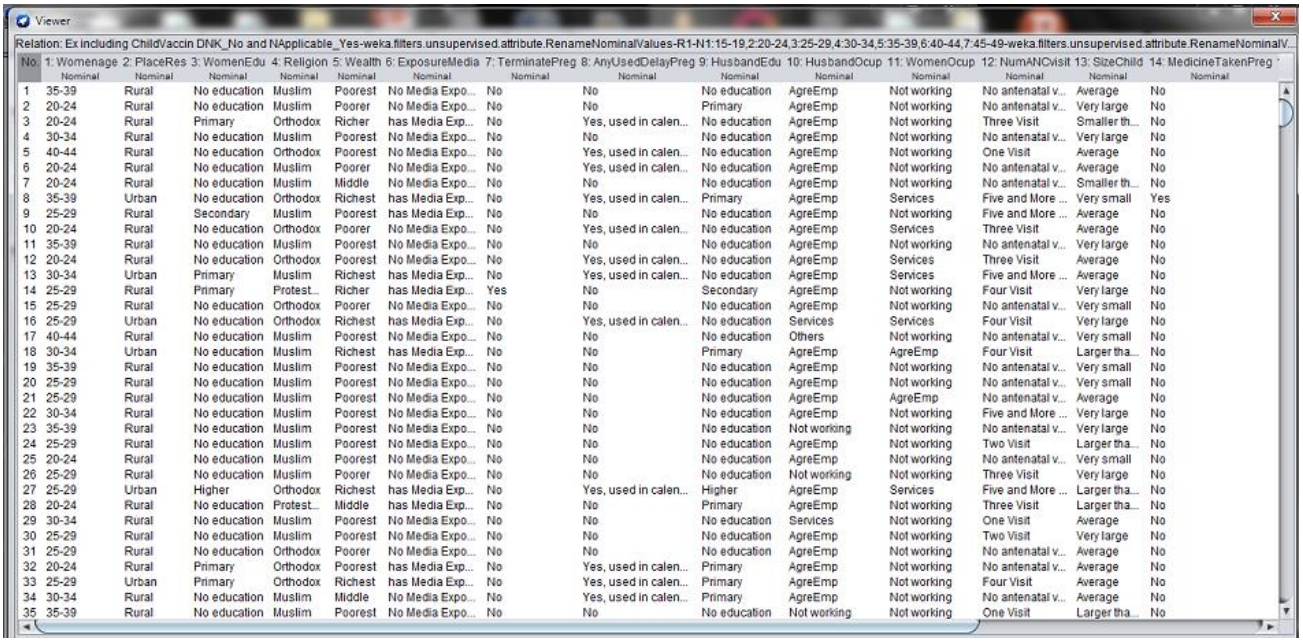
5.6. Classifier Error

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 designed/ trained to recognize. When the investigator talks 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 as 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 to use the model for prediction.

Sometimes, the predicted and actual value may differ in predicting a record to a certain class label. This shows that the record that 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 Random Forest algorithm with 80% split test option mode used 2,665 records for testing the model performance. 1,330 records and 1,317 records

correctly classified as "Alive" and "Dead" respectively. The classifier incorrectly classified 16 records as "Dead" and 2 as "Alive". The below Figure shows sample of instances that show the actual and predicted class difference.



No	1: Womenage	2: PlaceRes	3: WomenEdu	4: Religion	5: Wealth	6: ExposureMedia	7: TerminatePreg	8: AnyUsedDelayPreg	9: HusbandEdu	10: HusbandOcup	11: WomenOcup	12: NumANCvisit	13: SizeChild	14: MedicineTakenPreg
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	35-39	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Average	No
2	20-24	Rural	No education	Muslim	Poorer	No Media Expo...	No	No	Primary	AgreEmp	Not working	No antenatal v...	Very large	No
3	20-24	Rural	Primary	Orthodox	Richer	has Media Exp...	No	Yes, used in calen...	No education	AgreEmp	Not working	Three Visit	Smaller th...	No
4	30-34	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Very large	No
5	40-44	Rural	No education	Orthodox	Poorest	No Media Expo...	No	Yes, used in calen...	No education	AgreEmp	Not working	One Visit	Average	No
6	20-24	Rural	No education	Muslim	Poorer	No Media Expo...	No	Yes, used in calen...	No education	AgreEmp	Not working	No antenatal v...	Average	No
7	20-24	Rural	No education	Muslim	Middle	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Smaller th...	No
8	35-39	Urban	No education	Orthodox	Richest	has Media Exp...	No	Yes, used in calen...	Primary	AgreEmp	Services	Five and More ...	Very small	Yes
9	25-29	Rural	Secondary	Muslim	Poorest	has Media Exp...	No	No	No education	AgreEmp	Not working	Five and More ...	Average	No
10	20-24	Rural	No education	Orthodox	Poorer	No Media Expo...	No	Yes, used in calen...	No education	AgreEmp	Services	Three Visit	Average	No
11	35-39	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Very large	No
12	20-24	Rural	No education	Orthodox	Poorest	No Media Expo...	No	Yes, used in calen...	No education	AgreEmp	Services	Three Visit	Average	No
13	30-34	Urban	Primary	Muslim	Richest	has Media Exp...	No	Yes, used in calen...	No education	AgreEmp	Services	Five and More ...	Average	No
14	25-29	Rural	Primary	Protest...	Richer	has Media Exp...	Yes	No	Secondary	AgreEmp	Not working	Four Visit	Very large	No
15	25-29	Rural	No education	Orthodox	Poorer	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Very small	No
16	25-29	Urban	No education	Orthodox	Richest	has Media Exp...	No	Yes, used in calen...	No education	Services	Services	Four Visit	Very large	No
17	40-44	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	Others	Not working	No antenatal v...	Very small	No
18	30-34	Urban	No education	Muslim	Richest	has Media Exp...	No	No	Primary	AgreEmp	AgreEmp	Four Visit	Larger tha...	No
19	35-39	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Very small	No
20	25-29	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Very small	No
21	25-29	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	AgreEmp	No antenatal v...	Average	No
22	30-34	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	Five and More ...	Very large	No
23	35-39	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	Not working	Not working	No antenatal v...	Very large	No
24	25-29	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	Two Visit	Larger tha...	No
25	20-24	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Very small	No
26	25-29	Rural	No education	Muslim	Poorer	No Media Expo...	No	No	No education	Not working	Not working	Three Visit	Very large	No
27	25-29	Urban	Higher	Orthodox	Richest	has Media Exp...	No	Yes, used in calen...	Higher	AgreEmp	Services	Five and More ...	Larger tha...	No
28	20-24	Rural	No education	Protest...	Middle	has Media Exp...	No	No	Primary	AgreEmp	Not working	Three Visit	Larger tha...	No
29	30-34	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	Services	Not working	One Visit	Average	No
30	25-29	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	AgreEmp	Not working	Two Visit	Very large	No
31	25-29	Rural	No education	Orthodox	Poorer	No Media Expo...	No	No	No education	AgreEmp	Not working	No antenatal v...	Average	No
32	20-24	Rural	Primary	Orthodox	Poorest	has Media Exp...	No	Yes, used in calen...	Primary	AgreEmp	Not working	No antenatal v...	Average	No
33	25-29	Urban	Primary	Orthodox	Richest	has Media Exp...	No	Yes, used in calen...	Primary	AgreEmp	Not working	Four Visit	Average	No
34	30-34	Rural	No education	Muslim	Middle	No Media Expo...	No	Yes, used in calen...	Primary	AgreEmp	Not working	No antenatal v...	Average	No
35	35-39	Rural	No education	Muslim	Poorest	No Media Expo...	No	No	No education	Not working	Not working	One Visit	Larger tha...	No

Figure 5. 3: Sample of records that show the actual class and predicted class variation

As shown from the above Figure 5.3, the experts classified the pattern of under-five children survival status whether Alive or Dead based on EDHS 2016 database. Then the classifier wrongly classified the status of the under-five children. As it observed from the evaluation results on test data, the total error rate of this classifier was 2.71% where it wrongly classified 16 'Dead' cases as 'Alive' and 2 'Alive' cases as 'Dead'.

Consecutively, to see how well the predictive model can recognize "Alive" and how well the predictive model, which has the classifier, can recognize "Dead" which have sensitivity and specificity measures can 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 99% sensitivity and 98% specificity;

which discloses that the J48 decision tree classifier has an acceptable capability of recognizing the true class value. Hence, this indicated that records whose class is 'Alive' classified with a minimum error as compared with the records in the class 'Dead'.

5.7. Generating Rules from Random Forest Tree

Decision tree can produce a model with rules that are human-readable and interpretable[125]. The classification task using tree technique can be perform without complicated computations and the technique used for both continuous and categorical variables. This technique is suitable for predicting categorical outcomes[125] In present, there are many research that in use decision tree techniques and it is showed that, the decision tree is among the powerful classification algorithms some of decision tree classifiers are C4.5, J48, Random Forest, NBTree, SimpleCart, REPTree and others[126].

The C4.5 techniques is one of the decision tree families that can produce both decision tree and rule-sets; and construct a tree. Besides that, C4.5 models are easy to understand as the rules that derived from the technique have a very straightforward interpretation. A Random Forest is a classifier consisting of a collection of tree structured classifiers[88].

Random Forests[88] It constructs a series of classification trees which will be used to classify a new example. The idea used to create a classifier model is constructing multiple decision trees, each of which uses a subset of attributes randomly selected from the original set of attributes. However, the rules generated by existing ensemble techniques sometimes conflict with the rules generated from another classifier. This may lead to a problem when we want to combine all rule set into a single rule set. Therefore, several works intend to increase the accuracy of the classifiers[127].

A method for extracting rules from a decision tree is quite simple. A rule can be extract from a path linking from the root to a leaf node. All nodes in the path are gathered and connected to each other using conjunctive operations[127, 128]. This produces clear rules, even it does not matter whatever the order they executed. Decision tree and decision

rule solutions offer a level of interpretability that is unique to symbolic models. The complete set of decision rules generated by a decision tree is equivalent (for classification purposes) to the decision tree itself. The solutions may be directly inspected to understand the decision surfaces that exist in the data[129].

in this research when the researcher compared the performance measure as well as the results obtained from both algorithms (i.e. Random Forest and PART models) are Random forest is better performing than PART Rules; In terms of accuracy, precision, F-measure and AUC. Therefore, the researcher selected the Random Forest algorithm for generating better rules.

Decision rules come in the form if antecedent, then consequent, as shown below for each of the outcome variable. For decision rules, the antecedent consists of the attribute values from the branches taken by the particular path through the tree, while the consequent consists of the classification value for the target variable given by the particular leaf node. The following are some of the rules extracted from the Random Forest tree listed below and some of the rules that are interesting and selected by domain experts as well as from the literatures, presented as follows:

Rule: 1

Rule: 1.A

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** Mother place of Residence is Rural **AND** women are used Anything to delay pregnancy is Yes, used in calendar **AND** Women education status is Not Educated **AND** women has Drug exposure **AND** Women didn't conduct antenatal Visit **AND** child vaccinated is No **AND** the size of child is Average **THEN** likely of the under five children will be "**Dead**" (77/0)

Rule: 1.B

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** Mother place of Residence is Rural **AND** women are used Anything to delay pregnancy is Yes, used in calendar **AND** Women education status is Not Educated **AND** women has Drug exposure **AND** Women didn't conduct antenatal Visit

AND child vaccinated is No **AND** the size of child is very small **AND** women status of occupations is not working **THEN** likely of the under five children will be "**Dead**" (36/0)

These two Rules can minimize by removing the redundancy by Extend the range of continuous conditions. The rules with the range of the same attribute can be combined into the widest one according to Naphaporn Sirikulviriya et.al new Rules[127].

The New Rule are

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** Mother place of Residence is Rural **AND** women are used Anything to delay pregnancy is Yes, used in calendar **AND** Women education status is Not Educated **AND** women has Drug exposure **AND** Women didn't conduct antenatal Visit **AND** child vaccinated is No **AND** the size of child is very smaller **AND** women status of occupations is not working **THEN** likely of the under five children will be "**Dead**" (36/0)

Rule 2:

IF the Wealth status of the family is poorer **AND** Women ages is between 35-39 **AND** Mother place of Residence is Rural **AND** women are used Anything to delay pregnancy is Yes, used in calendar **AND** Women education status is Not Educated **AND** women has Drug exposure **AND** the size of child is very small **THEN** likely of the under five children will be "**Dead**" (32/0)

Rule 3:

IF the Wealth status of the family is poorer **AND** Women ages is between 45-49 **AND** Mother place of Residence is Rural **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Muslim **AND** child vaccinated is No **AND** Women didn't conduct antenatal Visit **AND** the women terminated pregnant is yes **THEN** likely of the under five children will be "**Dead**" (37/0)

Rule 4:

IF the Wealth status of the family is poorer **AND** Women ages is between 35-39 **AND** Mother place of Residence is Rural **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Muslim **AND** child vaccinated is No **AND** Women conduct antenatal Visit is five and more visit **AND** the size of child is very small **AND** Women education status is Not Educated **THEN** likely of the under five children will be "**Dead**" (48/0)

Rule 5:

IF the Wealth status of the family is poorer **AND** Women ages is between 35-39 **AND** Mother place of Residence is Rural **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Muslim **AND** child vaccinated is No **AND** Women conduct antenatal Visit is Four visit **AND** Women education status is Primary Educated **AND** Women put anything of child cord is Yes **THEN** likely of the under five children will be "**Dead**" (36/0)

Rule 6:

IF the Wealth status of the family is poorest **AND** Women ages is between 35-39 **AND** Mother place of Residence is Rural **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Muslim **AND** child vaccinated is No **AND** Women conduct antenatal Visit is No **AND** Husband occupation status are agriculture Employee **AND** the size of child is very large **AND** women are used Anything to delay pregnancy is No **THEN** likely of the under five children will be "**Dead**" (37/0)

Rule 7:

IF the Wealth status of the family is poorest **AND** Women ages is between 45-49 **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Muslim **AND** child vaccinated is No **AND** Women conduct antenatal Visit is five and more visit **AND** Husband occupation status are Not Working **THEN** likely of the under five children will be "**Dead**" (37/0)

Rule 8:

IF the Wealth status of the family is poorest **AND** Women ages is between 45-49 **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Orthodox **AND** Women status on Exposure media is No, Media Exposure **AND** the women terminated pregnant is Yes **AND** child vaccinated is No **AND** the size of child is very small **AND** Women has Drug Exposure **AND** women education status is Not Educated **THEN** likely of the under five children will be "**Dead**" (50/0)

Rule 9:

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Orthodox **AND** Women status on Exposure media is No, Media Exposure **AND** the women terminated pregnant is Yes **AND** child vaccinated is No **AND** the size of child is Very large **AND** women are used Anything to delay pregnancy is Yes **THEN** likely of the under five children will be "**Dead**" (38/0)

Rule 10:

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Orthodox **AND** Women status on Exposure media is No, Media Exposure **AND** the women terminated pregnant is Yes **AND** child vaccinated is No **AND** the size of child is Very large **AND** women are used Anything to delay pregnancy is Yes, used in calendar **AND** Women conduct antenatal Visit is Five and More visit **THEN** likely of the under five children will be "**Dead**" (40/0)

Rule 11:

IF the Wealth status of the family is poorest **AND** Women ages is between 35-39 **AND** women occupation status are not working **AND** Husband education status is Not Educated **AND** religious is Orthodox **AND** Women status on Exposure media is No, Media Exposure **AND** the women terminated pregnancy is No **AND** child vaccinated is No **AND** the size of child is Very Small **THEN** the likely of the under five children will be "**Dead**" (81/0)

Rule 12:

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** women occupation status are not working **AND** Husband education status is Primary **AND** religious is Muslim **AND** the size of child is Very small **AND** Women has No Drug Exposure **THEN** the likely of the under five children will be "**Dead**" (32/0)

Rule 13:

IF the Wealth status of the family is poorer **AND** Women ages is between 40-44 **AND** women occupation status are not working **AND** Husband education status is Primary **AND** religious is Orthodox **AND** the women terminated pregnant is Yes **AND** Women has Drug Exposure **AND** Women conduct antenatal Visit is Four visit **THEN** the likely of the under five children will be "**Dead**" (48/0)

Rule 14:

IF the Wealth status of the family is poorer **AND** Women ages is between 20-24 **AND** women occupation status are not working **AND** Husband education status is Primary **AND** religious is Orthodox **AND** the women terminated pregnant is Yes **AND** Women has No Drug Exposure **AND** Women status on Exposure media is No, Media Exposure **AND** Women education status is Primary **THEN** the likely of the under five children will be "**Dead**" (25/0)

Rule 15:

IF the Wealth status of the family is poorest **AND** Women ages is between 25-29 **AND** women occupation status are Agricultural Employee **AND** Husband education status is No Education **AND** Husband occupation status are Agricultural Employee **AND** religious is Muslim **AND** the size of child is Very Small **AND** women are used Anything to delay pregnancy is Yes **THEN** the likely of the under five children will be "**Dead**" (43/0)

Rule 16:

IF the Wealth status of the family is poorer **AND** Women ages is between 25-29 **AND** women occupation status are Agricultural Employee **AND** Husband education status is No Education **AND** Husband occupation status are Agricultural Employee **AND** religious is Orthodox **AND** the size of child is Average **AND** Women education status is No Education **AND** Women conduct antenatal Visit is No Antenatal visit **THEN** the likely of the under five children will be "**Dead**" (42/0)

Rule 17:

IF the Wealth status of the family is poorer **AND** Women ages is between 30-34 **AND** Women conduct antenatal Visit is No Antenatal visit **AND** women occupation status are Not Working **AND** the women terminated pregnancy is Yes **AND** the size of child is Larger than very small **AND** Women has Drug Exposure **AND** Husband education status is No Education **AND** child vaccinated is No **THEN** the likely of the under five children will be "**Dead**" (94/0)

5.8. Discussion on Generating Rules from the Classification Models

The following rules extracted by classification models. From the generated rules, it observed that the most determinant factors are Women Occupation, Place of Residence, Child Vaccine, Number of ANC Visit, anything applied on Cord, Husband education Wealth of the family, and Religion. When we see that (A woman occupation status) a non-working woman has chance the child to be die, even more worsen for not working girls living in rural village than urban living women. Rural women who are born 1-4 total children and No ANC visit and wealth level is poorer and not used anything to delay the pregnant and the child size is very small then most probably the child will be died if the women age between 35-39. The second case: A Rural women who are born 1-4 total children and No ANC visit and wealth level is poorest and the women age is 35-39 and husband not educated then the protestant then the child will died.

As we can see from the above rule sets: if every women conducting at list one antenatal visit is (one ANC) has the chance to save the child from the death because of most of the death occurred with women not visit the ANC at least once. This is more worsen when the women (the family) wealth ranked poorer or poorest the child death is higher in addition to not visiting ANC.

When comparing the child death by religious most of child death reported by Muslim, protestant and orthodox followers than traditional and other religious followers. When we discussed with some domain experts they assumed that some religious followers may not assume child death as human being deaths so they may not report it and the other

assumption could be; they may be believing that counting the death will be cases to lost the living children.

A woman who lived in rural village that have experienced the child may not vaccine properly so it shows that there is higher number of death occur through non-vaccine children (Like BCG, Tetanus, etc.). It could true as some domain experts justified this, because as they said, child vaccine medicine is very expensive and funded by donation. so there may be shortage so many of children may not get it and the other reason child vaccine is need special (like 24hr working refrigerators) so most of the rural village may not have those facility and electricity may not accessible. Therefore, most of the child may die due to not vaccine the essential medicine.

At all, some of the interesting rules show that ANC visit at least one times is so important, for all women whether they are at working or not it a must and their husband are educated at least primary will have the chance to reduce child death by guiding the women to health facility. The second things is child vaccine is mandatory for every child whether they born from poorer or poorest family and even they are live in rural or not so every facility should provide those vaccine easily. The third thing is every religious leader should tech their follower about child educating women, vaccine child, ANC visits and evil and non-evil things when reporting the truths thus those things increasing the chance of reducing the child mortality.

Chapter Six

Use of the Extracted Knowledge

6.1. Demonstrations of the Sample Test

The discovery of repeated patterns among data is the objective of data mining[59]. In particular, the extraction of classification rules, first proposed in Agrawal[130], is widely recognized as one of the most important paradigms. Most research work, [131-133] and also [134, 135], has been focused on the definition of efficient techniques to perform classification, association rules extraction with algorithms specifically designed in response to given user requests. To our knowledge, neither classification, nor design guidelines for rule extraction applications have been propos so far to guide the extraction process.

The importance of the definition of a unifying framework for classification and association rule mining is now recognized[136], and new languages are being proposed[137] to allow the specification of general extraction criteria. Powerful rule extraction languages, e.g. the query language operator in Meo[137], are oriented to expert SQL and Access programmers and may prove difficult to use for inexpert users, such as managers and executives, which exploit this type of analysis tools for business decisions.

In this paper, we propose a designing an interface for the classified rule extraction applications. We focus on the extraction of rules from database data (E.g. Access Database or SQL). We set the stage for the definition of the rule design problem, by identifying several classes of relevant classification rules, described in Agrawal and Han[135, 138] that can be extracted from the database data. Next, we evidence the common features of these classes, which allow us to pinpoint several extraction criteria, which are used to research relevant rules.

Based on the above criteria, we define a template language, which allows the specification of a predefined format for different extraction conditions, in which only the target

database and attributes must be instantiated. Hence, classification rule templates provide a simplified interface for defining rule extraction criteria. Inexpert users to extract interesting rules with a predefined structure can then use templates (e.g., sequential patterns) simply by instantiating database variables. Furthermore, the template definition language provides to expert rule programmers a flexible way to define new extraction criteria based on specifications provided by end-users. The template language described in this paper provides a theoretical foundation for a user-friendly interface. The definition of templates allows the dynamic generation of a graphic user-friendly interface, to provide a simple interactive way for specifying and instantiating templates.

The frontend of the user friendly interface looks like as Figure 6.1 which is designed by Visual Studio 2015 software (which is freely available in the web) to write the code. On this interface the variable name is used from the reduced variables so that for each variable here designed data entry box which is call (Text box) with labeled by variables name and follows the backend of the interfaces are used Microsoft Access/SQL database to store the data that received from the text box of variable interface. So the details of the access data base designing and the code linkage will explain below the picture of an interface of designing.

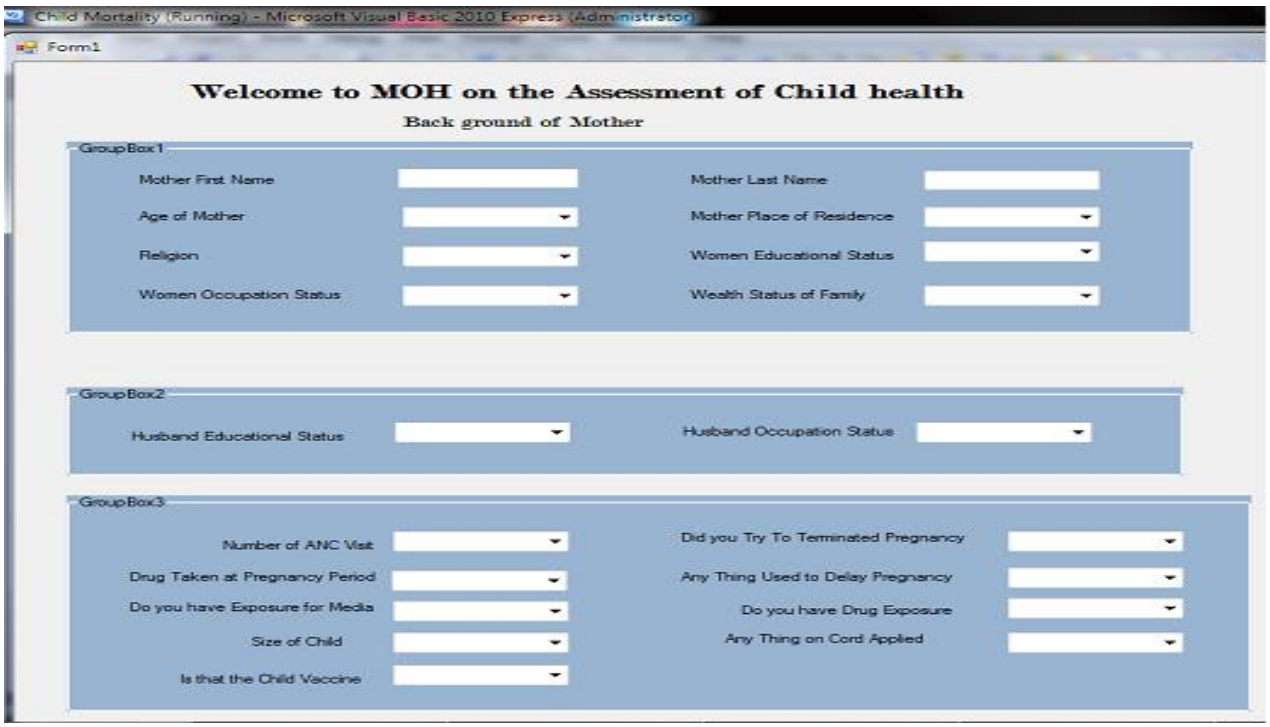


Figure 6. 1: Mother and Child registration Interface

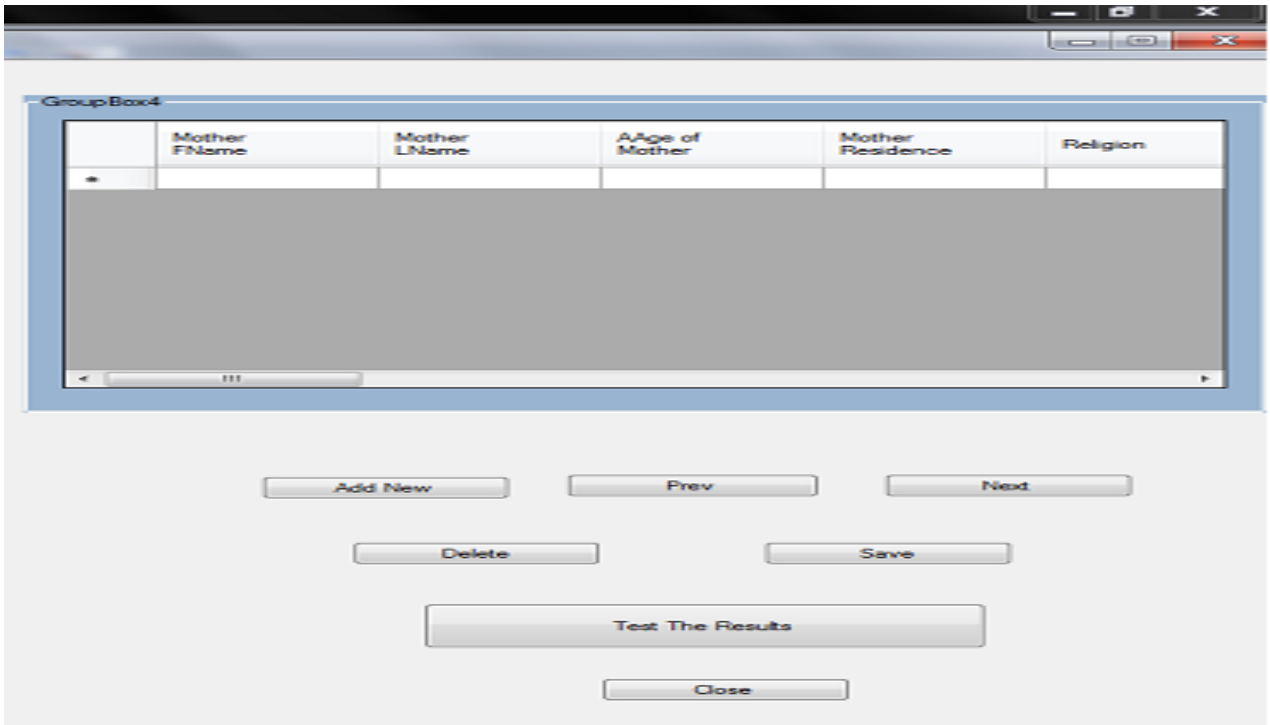


Figure 6. 2: database that Evaluate the status of Mother and Child

Let's describe the database relations on which will be based all the mining process in this paper. Therefore, they refer to the following relations:

AccessDatabase used as the name of the Database created on the Microsoft Access.
MotherInformation used as Table name, which have different variables.
The variables names are (Mother FName, Mother LName, Age of Mother, Mother Residence, Religion, Women Occupation, Women Education, WealthStatus of Family, Husband Educational Status, Husband Occupational Status, Number of ANC Visit, Terminated Pregnancy, Drug Taken at Pregnancy, Anything Used to Delay, Exposure for Media, Drug Exposure, Size of Child, Anything on Cord Applied, Child Vaccine)

AcessDatabase1Datasets used as Database Connectors with Visual basic Interface
MotherInformationBindingSources used as connecting variables with Data Grid interface on visual basics.
MotherInformationTableAdapter used as connecting the inputs from variable inputs from visual basic to Access database for add, Delete the data. The code that written inside of an interface is look like as in AppendixB.

In this study, the researcher currently defining a graphical user-friendly interface on top of using Microsoft Access/SQL language, to allow a simplified instantiation of variables for the chosen template. In particular, we are using the Visual Studio 2015 language to describe the interface, which customized on specific templates by means of Visual Basic scripts.

Therefore, in this paper we design an interface for the extraction of classification rule applications. We initially identified general classes of classification. Next, we described a template definition language, by means of which is possible to specify classes of classification rule extraction criteria. This language is defined as an extension of the SQL scripts[137], an SQL-like operator that allows the extraction of classification rules from relational databases. Templates provide a simplified interface for non-experienced users,

in which the extraction criterion may be predefined and only the target database must be instantiated. Templates for the most important extraction criteria shown, together with a complete description of the language features.

6.2. Evaluation of the Designed Interface

Evaluation is an essential activity in designing the interface for system development and implementation. Artifact evaluation should have seen as important part of the development process to make certain that the developed artifact can bring observed improvement and works in the real environment. Therefore, evaluation can be takes place during the designing of the Interface of prototype with the purpose of improving effectiveness and efficiency. The evaluation criteria focused on the issue of easy access and advantages of the interface platform. These criteria deal with the easy access of the interface platform to facilitate simplicity to member FMOH in order to gain the advantages from this platform.

6.2.1. Interface Testing

Testing is a technique used to perform to evaluate the complete system against specific requirements and it is an important and critical stage in software development. Testing plays an important role in the determining the quality and reliability of the application [22]. Different testing methods and tools used to ensure the functionality and usability of document management. The primary purpose of these tests is to uncover the systems limitations and measure its capabilities.

Following are testing tools used for all testing types:

A. Manual Testing

Manual test: - is a type of software testing where testers manually execute test cases without using any automation tools. Manual testing selected because it is the most primitive of all testing types and helps find bugs in the newly implemented system.

Testing made with different employees found in FMOH, such as MOM and Child health Case Teams office. These office users are the one who repeatedly interact with it. The key concept of having manual testing with those office users help; ensure that the implemented system is error free and it is working to the specified functional requirements.

The process of testing implemented with repetitive discussion between the system developer and system user. Finally, the end user fills the checklist form, which stated in each testing type. The checklist attached in Appendix B:

B. Unit Testing

In this testing module interface tested to assure that information properly and correctly flows into and out of the module. This testing involves the testing of data truncation, the structure of the data and the program correctly accepts the input data. The whole validation of the program encountered in this testing. Unit testing implemented and successfully tested, until it reaches a point where a set of methods are ready for the system user.

C. Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by Rule Creation so the goal of this testing is to verify the functionality of the proposed Rules.

Figure 6. 3: login Functionality with invalid inputs displayed error message

D. Usability Test

Usability testing is one of the most used methods to define the level of usability of a software product [139]. It intended to determine the extent an interface facilitates a user's ability to complete routine tasks. Those users are selected using purposive sampling methods because of most of the users are responsible for the mother and children health cases who working in FMOH especially for the mother and child department and sometimes for some circumstance some staffs may not avail when the interviews were conduct. That is why twenty users choose from FMOH employees avail at the time of interview. So, users asked to complete a series of routine tasks. Sessions recorded and analyzed to identify potential areas for improvement to the databased system.

Is the interface of the system user friendly?		Is the system attractive (regarding the font size and color combination) for use?		Is the system easy to understand and used?	
Yes	No	Yes	No	Yes	No
17	3	15	5	18	2
Does the elements or objects of the interface easily visible ?		Is the system allowing the system users to add additional meta data information about the record easily?		Is the system allowing the system user to search documents effectively?	
Yes	No	Yes	No	Yes	No
19	1	15	5	18	2

Table 6. 1: User Testing Result

As we have seen from the above usability test results, most of the respondents are satisfy by the designed interface and applicability of the systems. The interview done at FMOH office and the number of respondents was 20. Therefore, the respondents asked to complete the form related usability of the systems. The result of the questionnaires after the demonstration used to evaluate the interface and evaluated the results.

The average number of respondents with each of dimension was 18, meaning that; overall, the respondents were satisfied with the results of the interface. The values of interface user friendliness, system attractiveness, system easiness, easily visibility of interface object, and the system whether allowed to add or search or not with their satisfaction level are listed. Therefore, there result of the information are lists down in the Table 6.2 below.

Variables	Achievement of satisfaction	Achievement of non-satisfaction
Is the interface of the system user friendly	0.95	0.05
Is the system attractive (regarding the font size and color combination) for use	0.90	0.10
Is the system easy to understand and used?	0.85	0.15
Does the elements or objects of the interface easily visible ?	0.95	0.05
Is the system allowing the system users to add additional meta data information about the record easily?	0.80	0.20
Is the system allowing the system user to search documents effectively?	0.90	0.10

Table 6. 2: User Satisfaction on Usability of the interface

E. Finally

Testing result extracted from the manual testing made with different system testing participants. Repetitive interaction where made with each participants and finally, they give their last response using the checklist found in each testing types. Their checklist attached in the appendix section.

All testing participants are executing test cases manually and the following are agreed points:-

- The new system developed as their intended use.
- The system is easy to use.
- The system is well organized and helpful

Chapter Seven

Conclusion and Recommendations

7.1. Conclusion

Nowadays there are a number of improvements in machine learning intelligence with availability new software and algorithms added; this improvement helps healthcare data to be efficiently analyze because the data collected and stored by hiding valuable information. Even on, day to day the volume information is add so it makes difficult to get valuable information and to take action based on the data. Machine learning bring the new change on data mining and knowledge discovery from database. The importance of data mining increase time to time and its application area.

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 Ministry of Health specifically child health as well as for the governmental and non-governmental organizations to provide child health policy and strategy in Ethiopia.

This investigation, conducted according to the hybrid KDP model, 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 not followed strictly.

A dataset with 6,893 total EDHS 2016 records used to develop a classification model. Since, this research intended to fill a gap left by a related research; some valuable experiences of the previous research used. However, though the previous research's objective was to identify children mortality by using different datasets like Butajira Rural Health project (BRHP).

In the classification phase, the Random Forest tree algorithm and PART Rule classifier, which is WEKA's implementation algorithms used. From the study, around 18 attributes were choose for model building by applying different attribute selection methods and based on domain expert suggestion. In order to select a classification model that can classify the EDHS 2016 datasets, we used Random Forest (RF) tree algorithms. The researcher used Auto-Weka application and Experimental application methods to select Random Forest algorithms. Random Forest predicts better than PART Rules. In the process of selecting the tree the performance of accuracy and the size of tree generated has the more focused. However, all (18) attributes selected and used for the under-five mortality, for the current classification tasks.

The experiments were done using a WEKA version 3-8-2, it was observed that, for a given attribute, as the number of a random tree increases the corresponding number of rules generated possibly increase. Accordingly, better Random Forest with the corresponding rules selected as a working model to classify datasets into their corresponding classes. As a result, the classification accuracy of the selected Random Forest seems convincing than the PART Rule classifier. That is, among the 2,665 data inputted to the model learner with 80% split test mode, 99.32%, which is 2,647 records correctly classified.

The researcher was try to collect the domain expert suggestions and opinions throughout the entire investigation. So those comments and observation accepted because they are very important in the model development process, particularly, in the classification phase.

The overall predictive model building process made by using the Random Forest algorithm and PART rule classifier in the data mining methods predict the pattern of under-five mortality in Ethiopia, particularly for EDHS 2016 datasets and the results obtained from this study were interesting and encouraging; it can be used as decision support for healthcare policy maker. The extracted rules for each of the outcome variables are very effective for the prediction of child health. From the socio-economic and demographic variables used as predictor variables in this study for each of the outcome variables, it can be observed that the attributes such as Household Wealth, Place of

Residence, Women's Educational and Occupation level, Husbands Occupation and Educational level, Religion, Media exposure, and Drug Exposure are the most determinant factors to predict under-five child health.

7.2. Recommendations

This investigation has been conduct 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 EDHS 2016 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 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 such as Health facility/Hospital registry or Database, Regional specific survival studies..etc.
- 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.
- Here, in chapter five try to demonstrate the prototype using visual basic and MS-Access database, so further investigation could be add their future for further to elaborate functionality of the studies on under-fives children survival classification system.

- Further extensive experiments should require by using large amounts of dataset and applying different classification techniques.
- Investigations based on clinical datasets from different health facilities a need different DM research.
- There should be different DM research that can be undertake by comparing the DHS dataset with DSS dataset.
- Further study recommended to the problem domain specifically and under-fives children in general that apply those unused DM models, tools and algorithms.
- In this study, Random Forest and PART rules data mining techniques applied to predict under-five child health. However, more machine learning algorithms like Artificial Neural Network, Support Vector Machine and Multilayer Perception along with much larger data size needs to take to recognize the effects and optimize the prediction.
- The decision tree algorithm has achieved interesting results. Hence, an attempt should made to develop knowledge-based system (KBS) that would be helpful in assisting expert advice to identify the actual and non-actual user.

Reference

[1] W. H. Organization, *World Health Statistics 2016: Monitoring Health for the Sustainable Development Goals (SDGs)*. World Health Organization, 2016.

[2] U. N. D. o. Economic, S. Affairs, and P. Division, "World Population Prospects: The 2015 Revision, Key Findings and Advance Tables," *Working Paper, No. ESA/P/WP. 241.*, 2015.

[3] W. H. Organization, "WHO methods and data sources for country-level causes of death 2000–2012," *Geneva, Switzerland: WHO,* 2014.

[4] D. You, L. Hug, S. Ejdemyr, and J. Beise, "Levels and trends in child mortality. Report 2015. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation," 2015. UNICEF, 2015.

[5] U. N. C. Fund, *The State of the World's Children 2016: Executive Summary. A Fair Chance for Every Child*. UNICEF, 2016.

[6] Y. Mehretie Adinew, S. A. Feleke, Z. B. Mengesha, and S. B. Workie, "Childhood mortality: trends and determinants in Ethiopia from 1990 to 2015—A systematic review," *Advances in Public Health,* vol. 2017, pp. 1-10, 2017.

[7] U. Nations, "Sustainable Development Goals-17 goals to transform our world," ed: FAO, 2017, pp. 1-70.

[8] Y. Berhane, S. Wall, D. Kebede, A. Emmelin, and F. Enquselassie, "Establishing an epidemiological field laboratory in rural areas--potentials for public health research and interventions. The Butajira Rural Health Programme 1987-1999," 2000.

[9] M. Fantahun, "Mortality and survival from childhood to old age in rural Ethiopia," Folkhälsa och klinisk medicin, Umeå, vol. 17 PP. 1-62, 2008.

[10] E. M. Johnson, "The Handbook of Data Mining edited by Nong Ye 2003, 689 pages, $149.95 Mahwah NJ: Lawrence Erlbaum Associates ISBN 0-8058-4081-8," *Ergonomics in Design,* vol. 12, no. 3, pp. 27-27, 2004.

[11] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. Kurgan, *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.

[12] P. Ngom, F. N. Binka, J. F. Phillips, B. Pence, and B. Macleod, "Demographic surveillance and health equity in sub-Saharan Africa," *Health Policy and Planning,* vol. 16, no. 4, pp. 337-344, 2001.

[13] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

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

[15] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[16] C. S. A. C. E. a. ICF, " Ethiopia Demographic and Health Survey 2016: Key Indicators Report.," Reports October 2016 2016.

[17] T. Plate, P. Band, J. Bert, and J. Grace, "A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer," in *Advances in Neural Information Processing Systems*, 1997, pp. 967-973.

[18] W. H. Mosley and L. C. Chen, "An analytical framework for the study of child survival in developing countries," *Population and development review,* vol. 10, no. 0, pp. 25-45, 1984.

[19] Wikipedia contributors, "Infant mortality," (in English), 21 September 2017. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Infant_mortality&oldid=800403183. Wikipedia, The Free Encyclopedia.
[20] Federal Democratic Republic of Ethiopia Ministry, "HSDP IV Annual Performance Report," (in English), vol. Version 1, EFY 2006 (2013/14).
[21] N. P. C. Federal Democratic Republic of Ethiopia, "Growth and Transformation Plan II (GTP II) (2015/16-2019/20)," (in English), Plan vol. Volume I : Main Text, May, 2016 2016.
[22] K. A. Kyei, "Socio–economic factors affecting under five mortality in South Africa–An investigative study," *Journal of Emerging Trends in Economics and Management Sciences*, vol. 2, no. 2, pp. 104-110, 2011.
[23] L. Huicho, C. A. Huayanay-Espinoza, P. Hernandez, J. N. de Guzman, and M. Rivera-Ch, "Enabling reproductive, maternal, neonatal and child health interventions: Time trends and driving factors of health expenditure in the successful story of Peru," *PloS one*, vol. 13, no. 10, p. e0206455, 2018.
[24] R. E. Black *et al.*, "Reproductive, maternal, newborn, and child health: key messages from Disease Control Priorities 3rd Edition," *The Lancet*, vol. 388, no. 10061, pp. 2811-2824, 2016.
[25] M. Mahy, *Childhood mortality in the developing world: a review of evidence from the Demographic and Health Surveys*. MEASURE DHS+, ORC Macro, 2003.
[26] K. Abera, "Retrospective cohort study in the determinants of child mortality in BRHP and DSS," Thesis 2006. Addis Ababa University, Ethiopia.
[27] T. A. Houweling and A. E. Kunst, "Socio-economic inequalities in childhood mortality in low- and middle-income countries: a review of the international evidence," *British medical bulletin*, vol. 93, no. 1, pp. 7-26, 2009.
[28] S. El-Saharty, S. Kebede, P. Olango Dubusho, and B. Siadat, "Ethiopia: improving health service delivery," 2009.
[29] A. Shegaw, "Application of data mining technology to predict child mortality patterns: the case of Butajira Rural Health Project (BRHP)," aau, 2002.
[30] Amanuel, "Application of Data Mining Techniques to Predict Household Health Seeking Patterns: The Case of BRHP.," Thesis 2011. Addis Ababa University; Ethiopia.
[31] B. Taddesse, "Mining Vital Statistics Data: The case of BRHP," MSc. Thesis, Addis Ababa University, Ethiopia, 2011.
[32] W. H. Organization, *Health in 2015: From MDGs, millennium development goals to SDGs, sustainable development goals*. World Health Organization, 2015.
[33] World Health Organization (WHO), "children: Reducing Mortality fact Sheet,," (in English), Reducing Mortality 2017 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs178/en/ WHO.
[34] United Nations International Children's Emergency Fund (UNICEF), "Goal: Reduce Child Mortality," (in English), Web Site 21 September 2017. [Online]. Available: https://www.unicef.org/mdg/childmortality.html.
[35] S. Jāhāna, *Human development report 2015: Work for human development*. United Nations Development Programme, 2015.
[36] J. Tulloch, "Integrated approach to child health in developing countries," *The Lancet*, vol. 354, pp. SII16-SII20, 1999.
[37] W. H. Organization, *World health statistics 2015*. World Health Organization, 2015.

[38] M. J. Berry and G. Linoff, *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.
[39] Z. Markos, "Predicting Under nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demographic and Health Survey," Addis Ababa University, 2013.
[40] S. Laxman and P. S. Sastry, "A survey of temporal data mining," *Sadhana,* vol. 31, no. 2, pp. 173-198, 2006.
[41] A. Milley, "Healthcare and data mining," *Health Management Technology,* vol. 21, no. 8, pp. 44-45, 2000.
[42] M. Taft *et al.*, "Oracle Data Mining Concepts, 10g Release 2 (10.2) B14339-01," 2005.
[43] C. Oprean, "Towards user assistance in Data Mining," MSc. Thesis, University of Waterloo, Ontario, Canada, 2011.
[44] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*. AAAI press Menlo Park, 1996.
[45] G. Piateski and W. Frawley, *Knowledge discovery in databases*. MIT press, 1991.
[46] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma, "A knowledge discovery approach to diagnosing myocardial perfusion," *IEEE Engineering in Medicine and Biology Magazine,* vol. 19, no. 4, pp. 17-25, 2000.
[47] P. Julio and K. Adem, *Data mining and knowledge discovery in real life applications*. 2011.
[48] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 17, no. 3, p. 37, 1996.
[49] G. Piatetsky-Shapiro, "The data-mining industry coming of age," *IEEE Intelligent Systems and their Applications,* vol. 14, no. 6, pp. 32-34, 1999.
[50] A. I. R. L. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," *IADS-DM,* 2008.
[51] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *Journal of data warehousing,* vol. 5, no. 4, pp. 13-22, 2000.
[52] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step data mining guide," 2000.
[53] J. Jaklič, "The deployment of data mining into operational business processes," in *Data mining and knowledge discovery in real life applications*: InTech, 2009.
[54] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.
[55] S. Sumathi and S. Sivanandam, *Introduction to data mining and its applications*. Springer, 2006.
[56] M. Theeuwen, H. Kappen, and J. Neijt, "Neural network analysis to predict treatment outcome in patients with ovarian cancer," in *Industrial Applications of Neural Networks*: World Scientific, 1998, pp. 433-438.
[57] J.-s. Li, H.-y. Yu, and X.-g. Zhang, "Data Mining in Hospital Information System," in *New Fundamental Technologies in Data Mining*: InTech, 2011.
[58] S. Deshpande and V. Thakare, "Data mining system and applications: A review," *International Journal of Distributed and Parallel systems (IJDPS),* vol. 1, no. 1, pp. 32-44, 2010.
[59] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM,* vol. 39, no. 11, pp. 27-34, 1996.

[60] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," in *KDD*, 1996, vol. 96, pp. 82-88.
[61] M. A. Bramer, "Principles of Data Mining, ser. Undergraduate Topics in Computer Science," ed: London, UK: Springer, 2007.
[62] S. Velickov and D. Solomatine, "Predictive data mining: practical examples," in *2nd Joint Workshop on Applied AI in Civil Engineering*, 2000.
[63] S. M. Weiss and C. A. Kulikowski, *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., 1991.
[64] D. J. Hand, "Discrimination and classification," *Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley, 1981*, 1981.
[65] T. C. Corporation, *Introduction to data mining and knowledge discovery*. Two Crows Corporation, 1999.
[66] L. Breiman, *Classification and regression trees*. Routledge, 2017.
[67] S. Deolekar and S. Abraham, "Tree based classification of tabla strokes," *arXiv preprint arXiv:1801.01712*, 2018.
[68] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197-227, 2016.
[69] D. Getu and T. Fasil, "Biostatistics lecture note for health sciences students," *The Carter Center, Addis Ababa, Ethiopia*, vol. 181, 2005.
[70] M. W. Berry and M. Browne, *Lecture notes in data mining*. World Scientific, 2006.
[71] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. MIT press, 2001.
[72] M. Lloyd-Williams, "Discovering the hidden secrets in your data-the data mining approach to information," *Information research*, vol. 3, no. 2, 1997.
[73] S. Vinterbo, "Predictive models in medicine: some methods for construction and adaptation," *Department of Computer and Information Science. Norwegian University of Science and Technology*, 1999.
[74] S. Mitra and T. Acharya, *Data mining: multimedia, soft computing, and bioinformatics*. John Wiley & Sons, 2005.
[75] J. Ranjan, "Data mining in pharma sector: benefits," *International journal of health care quality assurance*, vol. 22, no. 1, pp. 82-92, 2009.
[76] M. K. Obenshain, "Application of data mining techniques to healthcare data," *Infection Control & Hospital Epidemiology*, vol. 25, no. 8, pp. 690-695, 2004.
[77] P. Cerrito, *Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks: Studies and Frameworks*. IGI Global, 2010.
[78] H. Kaur and S. K. Wasan, "Empirical study on applications of data mining techniques in healthcare," *Journal of Computer science*, vol. 2, no. 2, pp. 194-200, 2006.
[79] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of healthcare information management*, vol. 19, no. 2, p. 65, 2011.
[80] T. K. Helen, "Application of data mining technology to identify significant patterns in census or survey data: the case of 2001 child labor survey in Ethiopia," aau, 2003.
[81] D. AYELE, "PREDICTING MATERNAL HEALTH CARE SEEKING PATTERN USING DATA MINING TECHNIQUES IN ETHIOPIA " June 2013. Addis Ababa University.
[82] T. Hailemariam, "Application of data mining for predicting adult mortality," *Master's thesis. Addis Ababa, Ethiopia: Addis Ababa University*, 2012.

[83] N. Levin and J. Zahavi, "Data mining for target marketing," in *Data Mining and Knowledge Discovery Handbook*: Springer, 2005, pp. 1261-1301.
[84] M. Sharma and M. K. Borana, "Clustering in data mining: A brief review," *International Journal Of Core Engineering & Management (IJCEM),* 2014.
[85] E. Frank, L. Trigg, G. Holmes, and I. H. Witten, "Naive Bayes for regression," *Machine Learning,* vol. 41, no. 1, pp. 5-25, 2000.
[86] H. Sug, "Applying randomness effectively based on random forests for classification task of datasets of insufficient information," *Journal of Applied Mathematics,* vol. 2012, 2012.
[87] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 2012: Springer, pp. 154-168.
[88] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.
[89] D. Steinberg, M. Golovnya, and N. S. Cardell, "Data Mining with Random Forests™," 2004.
[90] L. Breiman, "Random forests leo breiman and adele cutler," *Random Forests-Classification Description*, 2015.
[91] H. Sug, "An effective sampling method for decision trees considering comprehensibility and accuracy," *W. Trans. on Comp,* vol. 8, no. 4, pp. 631-640, 2009.
[92] A. Mahajan and A. Ganpati, "Performance evaluation of rule based classification algorithms," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol,* vol. 3, pp. 3546-3550, 2014.
[93] B. R. Gaines and P. Compton, "Induction of ripple-down rules applied to modeling large databases," *Journal of Intelligent Information Systems,* vol. 5, no. 3, pp. 211-228, 1995.
[94] S. Ali and K. A. Smith, "On learning algorithm selection for classification," *Applied Soft Computing,* vol. 6, no. 2, pp. 119-138, 2006.
[95] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
[96] A. Mahajan and A. Ganpati, "Performance evaluation of rule based classification algorithms," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* vol. 3, no. 10, pp. 3546-3550, 2014.
[97] M. Thangaraj and C. Vijayalakshmi, "Performance Study on Rule-based Classification Techniques across Multiple Database Relations," *International Journal of Applied Information Systems,* vol. 5, no. 4, pp. 1-7, 2013.
[98] M. Kantardzic, *Data Reduction*. Wiley Online Library, 2003.
[99] A. Selam, "Predicting the Occurrence of Measles Outbreak in Ethiopia Using DM Technology," MSc. Thesis, Addis Ababa University, Ethiopia, 2011.
[100] A. Fallahi and S. Jafari, "An expert system for detection of breast cancer using data preprocessing and bayesian network," *International Journal of Advanced Science and Technology,* vol. 34, pp. 65-70, 2011.
[101] S. L. Braver, D. P. MacKinnon, and M. Page, *Levine's guide to SPSS for analysis of variance*. Psychology Press, 2003.
[102] D. T. Larose, *Data mining methods & models*. John Wiley & Sons, 2006.
[103] E. N. Ogor, "Student academic performance monitoring and evaluation using data mining techniques," in *Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007*, 2007: IEEE, pp. 354-359.

[104] W. Huang, "Economic Growth and Under Five Mortality-A systematic review of literature," *Master of Science thesis submitted in International Social Welfare and Health Policy. Faculty of Social Sciences, Oslo University College*, 2009.
[105] T. F. Asena, D. Dagne, and A. S. Bedane, "Determinats of Child Mortality in Arba Minch Hospital," *International Journal of Healthcare and Medical Sciences,* vol. 1, no. 3, pp. 27-35, 2015.
[106] M. Uddin, "Child Mortality in a Developing Country: A Statistical Analysis," *Journal of Applied Quantitative Methods,* vol. 4, no. 3, pp. 270-283, 2009.
[107] S. Piramuthu, "Evaluating feature selection methods for learning in data mining applications," *European journal of operational research,* vol. 156, no. 2, pp. 483-494, 2004.
[108] I. L. Organization, "International Standards Classification of Occupations (ISCO)," 2004. [Online]. Available: https://www.ilo.org/public/english/bureau/stat/isco/isco88/major.htm.
[109] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707*, 2013.
[110] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge & Data Engineering*, no. 9, pp. 1263-1284, 2008.
[111] T. Oommen, L. G. Baise, and R. M. Vogel, "Sampling bias and class imbalance in maximum-likelihood logistic regression," *Mathematical Geosciences,* vol. 43, no. 1, pp. 99-120, 2011.
[112] M. Anis, M. Ali, and A. Yadav, "A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection," *International Journal of Economics, Commerce and Management*, vol. 3, no. 12, 2015.
[113] D. Newby, A. A. Freitas, and T. Ghafourian, "Coping with unbalanced class data sets in oral absorption models," *Journal of chemical information and modeling*, vol. 53, no. 2, pp. 461-474, 2013.
[114] J. M. Choi, "A selective sampling method for imbalanced data learning on support vector machines," 2010.
[115] A. Sonak and R. Patankar, "A survey on methods to handle imbalance dataset," *Int. J. Comput. Sci. Mobile Comput,* vol. 4, no. 11, pp. 338-343, 2015.
[116] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, J. Ye, and A. s. D. N. Initiative, "Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study," *NeuroImage,* vol. 87, pp. 220-241, 2014.
[117] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classification: Algorithms and Applications,* p. 37, 2014.
[118] B. Jantawan and C.-F. Tsai, "A comparison of filter and wrapper approaches with data mining techniques for categorical variables selection," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 6, pp. 4501-4508, 2014.
[119] A. G. Karegowda, M. Jayaram, and A. Manjunath, "Feature subset selection problem using wrapper approach in supervised learning," *International journal of Computer applications,* vol. 1, no. 7, pp. 13-17, 2010.
[120] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *J. Mach. Learn. Res,* vol. 15, no. 1, pp. 3133-3181, 2014.
[121] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications,* vol. 39, no. 3, pp. 3446-3453, 2012.

[122] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton-Brown, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA," *The Journal of Machine Learning Research,* vol. 18, no. 1, pp. 826-830, 2017.
[123] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems*, 2015, pp. 2962-2970.
[124] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation,* vol. 1, no. 1, pp. 67-82, 1997.
[125] A. Rajput, R. P. Aharwal, M. Dubey, S. Saxena, and M. Raghuvanshi, "J48 and JRIP rules for e-governance data," *International Journal of Computer Science and Security (IJCSS),* vol. 5, no. 2, p. 201, 2011.
[126] J. Mesarić and D. Šebalj, "Decision trees for predicting the academic success of students," *Croatian Operational Research Review,* vol. 7, no. 2, pp. 367-388, 2016.
[127] N. Sirikulviriya and S. Sinthupinyo, "Integration of rules from a random forest," in *International Conference on Information and Electronics Engineering*, 2011, vol. 6, pp. 194-198.
[128] H. T. Bao, "Knowledge discovery and data mining techniques and practice," *Last Accessed,* vol. 28, 2005.
[129] C. Apté and S. Weiss, "Data mining with decision trees and decision rules," *Future generation computer systems,* vol. 13, no. 2-3, pp. 197-210, 1997.
[130] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, 1993, vol. 22, no. 2: ACM, pp. 207-216.
[131] R. Agrawal and R. Srikant, "Mining sequential patterns," in *icde*, 1995, vol. 95, pp. 3-14.
[132] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, 1994, vol. 1215, pp. 487-499.
[133] A. Savasere, E. R. Omiecinski, and S. B. Navathe, "An efficient algorithm for mining association rules in large databases," Georgia Institute of Technology, 1995.
[134] H. Toivonen, "Sampling large databases for association rules," in *VLDB*, 1996, vol. 96, pp. 134-145.
[135] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307-328, 1996.
[136] T. Imielinski and H. Mannila, "A database perspective on knowledge discovery," *Communications of the ACM,* vol. 39, no. 11, pp. 58-64, 1996.
[137] R. Meo, G. Psaila, and S. Ceri, "A new SQL-like operator for mining association rules," in *VLDB*, 1996, vol. 96, pp. 122-133.
[138] J. Han, Y. Fu, and S. Tang, "Advances of the DBLearn system for knowledge discovery in large databases," in *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, 1995: Morgan Kaufmann Publishers Inc., pp. 2049-2050.
[139] F.Paz, D. Villanueva, and J. Pow-sang "Heuristic Evaluation as a Complement to Usability Testing: A Case Study in Web Domain," in *in International Conference on Information Technology - New Generations (ITNG)*, Las Vegas, 2015.

A. Appendix (Participant Response)

- **Participant's Response Questioners attachment**

A. Unit Tests

Testing Activity	Yes	No	Remark
The first page appears on the very first call to a webpage.			
The system notifies the user if he tries to login with only user id or only password.			
System text fields and buttons aligned properly.			
The system forms display text fields in an order.			
Links for 'previous' and 'next' pages works fine the system.			
The system pages navigate according to a proper sequence.			
The system fonts and background colors are soothing to the user's eyes.			

B. Functional Tests

Testing Activity	Yes	No	Remark
All the links of interface are working correctly and successful redirect to another page.			
All Forms are working as expected and if user doesn't feel a mandatory field in form an error message is shown.			
All pages of the applications are displayed and working fine.			
Login functionality is working with valid inputs.			
Login functionality with invalid inputs displayed error message.			
All data manipulation is working like: Delete/Edit operations.			
All elements or objects of the Interface is working like: Buttons			
Is the system allowing the system users to add additional data information?			
Does the system allow the users to store records based on the classification schemes?			

C. Usability Tests

Testing Activity	Yes	No	Remark
Is the interface of the system user friendly?			
Is the system attractive (regarding the font size and color combination) for use?			
Is the system easy to understand and to be used?			
Does the elements or objects of the Interface is easily visible?			
Is the system allowing the system user to search documents effectively?			

D.Anexed (Code of the Interface)

```
Public Class Form1


    Private Sub Form1_Load(sender As System.Object, e As System.EventArgs) Handles
MyBase.Load
        LoginForm2.Show()
        Me.Hide()

        'TODO: This line of code loads data into the
'AccessDatabase1DataSet2.MotherInformation' table. You can move, or remove it, as needed.
        Me.MotherInformationTableAdapter.Fill(Me.AccessDatabase1DataSet2.MotherInformation)
        'TODO: This line of code loads data into the
'AccessDatabase1DataSet.MotherInformation' table. You can move, or remove it, as needed.
        Me.MotherInformationTableAdapter.Fill(Me.AccessDatabase1DataSet2.MotherInformation)
        'TODO: This line of code loads data into the
'AccessDatabase1DataSet1.MotherInformation' table. You can move, or remove it, as needed.
        Me.MotherInformationTableAdapter.Fill(Me.AccessDatabase1DataSet2.MotherInformation)
        'TODO: This line of code loads data into the
'AccessDatabase1DataSet.MotherInformation' table. You can move, or remove it, as needed.
        Me.MotherInformationTableAdapter.Fill(Me.AccessDatabase1DataSet2.MotherInformation)

        Me.Hide()
        LoginForm2.Show()
 End Sub

    Private Sub btnAddNew_Click(sender As System.Object, e As System.EventArgs) Handles
btnAddNew.Click


MotherInformationBindingSource.AddNew()
        Mother_FName.Focus()
        btnDelet.Enabled = False
        btnClose.Enabled = False
        btnNext.Enabled = False
        btnPrevious.Enabled = False
        btnSave.Enabled = False
        btnTestResult.Enabled = False
   End Sub

    Private Sub btnPrevious_Click(sender As System.Object, e As System.EventArgs) Handles
btnPrevious.Click
        MotherInformationBindingSource.MovePrevious()
    End Sub

    Private Sub btnNext_Click(sender As System.Object, e As System.EventArgs) Handles
btnNext.Click
        MotherInformationBindingSource.MoveNext()
    End Sub

    Private Sub btnDelet_Click(sender As System.Object, e As System.EventArgs) Handles
btnDelet.Click
        MotherInformationBindingSource.RemoveCurrent()
    End Sub
```

```
    Private Sub btnSave_Click(sender As System.Object, e As System.EventArgs) Handles
btnSave.Click
        On Error GoTo SaveErr
        MotherInformationBindingSource.EndEdit()
        MotherInformationTableAdapter.Update(AccessDatabase1DataSet2.MotherInformation)
        MessageBox.Show("ok Boss Successfully Saved  ")
SaveErr:
        Exit Sub
    End Sub

    Private Sub btnTestResult_Click(sender As System.Object, e As System.EventArgs) Handles
btnTestResult.Click
        Dim i As New Integer   , Dim x As Integer   , Dim y As Integer   , Dim A As Boolean
Dim B As Boolean   , Dim C As Boolean   , Dim D As Boolean   , Dim F As Boolean
Dim M As String  ,  Dim N As String , Dim K As String , Dim L As String , Dim H As String
Dim J As String  ,  Dim O As String , Dim P As String , Dim Q As String , Dim R As String
Dim S As String   , Dim T As String   , Dim U As Boolean   , Dim W As Boolean   , Dim V As
String
Dim Z As String   , Dim AB As Boolean   , Dim AC As String   , Dim AD As String
Dim AE As Boolean   , Dim AF As String   , Dim AG As String   , Dim AH As Boolean   , Dim AI
As String
Dim AJ As String   , Dim AK As Boolean   , Dim AL As String   , Dim AM As String
Dim AN As Boolean   , Dim AO As String   , Dim AP As String   , Dim AQ As Boolean   ,
Dim AR As String   , Dim AT As String   , Dim AU As Boolean   , Dim AV As String   ,
Dim AW As String   , Dim AX As Boolean   , Dim AY As String   , Dim AZ As String   ,

N = CboPlaceRes.Items(1)
        For y = 0 To CboPlaceRes.Items.Count
            Try
                M = CboPlaceRes.SelectedItem.ToString()

                F = (String.Equals(M, N))
                If (F = True) Then
                    'MessageBox.Show("    =output of selected index when F = " & F & " = " &
"String of Equal results /CboPlaceRes.SelectedIndex.ToString/ =" &
CboPlaceRes.SelectedIndex.ToString(0) & "  and  " & CboPlaceRes.SelectedItem.ToString())
                    Exit For
                End If
            Catch ex As Exception

            End Try
        Next

        Dim ABC As Boolean
        Dim ABA As String
        Dim ABB As Boolean

        J = CboAgeMother.Items(4)
        ABA = CboAgeMother.Items(5)
        L = CboAgeMother.Items(6)

        For i = 0 To CboAgeMother.Items.Count
            Try
                H = CboAgeMother.SelectedItem.ToString()
```

```
                C = (String.Equals(H, J))
                ABB = (String.Equals(H, ABA))
                ABC = (String.Equals(H, L))

                If (C = True) Then
                    'MessageBox.Show("    =output of selected index when C =" & C & " = " &
"String of Equal results /CboAgeMother.SelectedIndex.ToString/ =" &
CboAgeMother.SelectedIndex.ToString(0) & "  and   " & CboAgeMother.SelectedItem.ToString())
                    ABC = C
                    ABB = C
                    Exit For
                ElseIf (ABB = True) Then
                    ' MessageBox.Show("    =output of selected index when ABC =" & ABB & " =
" & "String of Equal results /CboAgeMother.SelectedIndex.ToString/ =" &
CboAgeMother.SelectedIndex.ToString(0) & "  and   " & CboAgeMother.SelectedItem.ToString())
                    C = ABB
                    ABC = ABB
                    Exit For
                ElseIf (ABC = True) Then
                    ' MessageBox.Show("    =output of selected index when ABC =" & ABC & " =
" & "String of Equal results /CboAgeMother.SelectedIndex.ToString/ =" &
CboAgeMother.SelectedIndex.ToString(0) & "  and   " & CboAgeMother.SelectedItem.ToString())
                    C = ABC
                    ABB = ABC
                    Exit For
                End If
            Catch ex As Exception

            End Try
        Next

        Dim ABE As String
        Dim ABF As Boolean
        P = CboWLFamily.Items(0)
        ABE = CboWLFamily.Items(1)

        For x = 0 To CboWLFamily.Items.Count
            Try
                O = CboWLFamily.SelectedItem.ToString()

                A = (String.Equals(O, P))
                ABF = (String.Equals(O, ABE))
                If (A = True) Then
                    ' MessageBox.Show("    =output of selected index when A =" & A & " = " &
"String of Equal results /CboWLFamily.SelectedIndex.ToString/ =" &
CboWLFamily.SelectedIndex.ToString(0) & "  and " & CboWLFamily.SelectedItem.ToString())
                    ABF = A
                    Exit For
                ElseIf (ABF = True) Then
                    ' MessageBox.Show("    =output of selected index when ABF =" & ABF & " =
" & "String of Equal results /CboWLFamily.SelectedIndex.ToString/ =" &
CboWLFamily.SelectedIndex.ToString(0) & "  and " & CboWLFamily.SelectedItem.ToString())
                    A = ABF
                    Exit For
                End If
            Catch ex As Exception
```

```
            End Try
        Next

        Dim Q As String
        Dim R As Boolean

        T = CboWmOccup.Items(0)
        Q = CboWmOccup.Items(1)

        For x = 0 To CboWmOccup.Items.Count
            Try
                S = CboWmOccup.SelectedItem.ToString()

                B = (String.Equals(S, T))
                R = (String.Equals(S, Q))
                If (B = True) Then
                    ' MessageBox.Show("    =output of selected index when B =" & B & " = " &
"String of Equal results /CboWmOccup.SelectedIndex.ToString/ =" &
CboWmOccup.SelectedIndex.ToString(0) & "  and " & CboWmOccup.SelectedItem.ToString())
                    R = B
                    Exit For
                ElseIf (R = True) Then
                    ' MessageBox.Show("    =output of selected index when R =" & R & " = " &
"String of Equal results /CboWmOccup.SelectedIndex.ToString/ =" &
CboWmOccup.SelectedIndex.ToString(0) & "  and " & CboWmOccup.SelectedItem.ToString())
                    B = R
                    Exit For
                End If
            Catch ex As Exception
            End Try
        Next

        Dim ABL As String
        Dim ABM As String
        Dim ABN As String
        Dim ABO As Boolean
        Dim ABP As Boolean

        ABL = CboHusOccup.Items(0)
        ABM = CboHusOccup.Items(1)

        For x = 0 To CboHusOccup.Items.Count

            Try
                ABN = CboHusOccup.SelectedItem.ToString()

                ABO = (String.Equals(ABN, ABL))
                ABP = (String.Equals(ABN, ABM))
                If (ABO = True) Then
                    ' MessageBox.Show("    =output of selected index when ABO =" & ABO & " =
" & "String of Equal results /CboHusOccup.SelectedIndex.ToString/ =" &
CboHusOccup.SelectedIndex.ToString(0) & "  and " & CboHusOccup.SelectedItem.ToString())
                    ABP = ABO
                    Exit For
                ElseIf (ABP = True) Then
```

```
                    ' MessageBox.Show("    =output of selected index when ABP =" & ABP & " =
" & "String of Equal results /CboHusOccup.SelectedIndex.ToString/ =" &
CboHusOccup.SelectedIndex.ToString(0) & "  and " & CboHusOccup.SelectedItem.ToString())
                    ABO = ABP
                    Exit For
                End If
            Catch ex As Exception
                Exit For
            End Try

        Next

        Dim U As Boolean
        Dim ABD As String
        Z = CboWmEdu.Items(0)
        ABD = CboWmEdu.Items(2)

        For x = 0 To CboWmEdu.Items.Count
            Try
                V = CboWmEdu.SelectedItem.ToString()

                W = (String.Equals(V, Z))
                U = (String.Equals(V, ABD))
                If (W = True) Then
                    ' MessageBox.Show("    =output of selected index when W =" & W & " = " &
"String of Equal results /CboWmEdu.SelectedIndex.ToString/ =" &
CboWmEdu.SelectedIndex.ToString(0) & "  and " & CboWmEdu.SelectedItem.ToString())
                    U = W
                    Exit For
                ElseIf (U = True) Then
                    ' MessageBox.Show("    =output of selected index when U =" & U & " = " &
"String of Equal results /CboWmEdu.SelectedIndex.ToString/ =" &
CboWmEdu.SelectedIndex.ToString(0) & "  and " & CboWmEdu.SelectedItem.ToString())
                    W = U
                    Exit For
                End If
            Catch ex As Exception
            End Try
        Next

        Dim K As Boolean
        Dim ABI As String
        Dim ABJ As String
        Dim ABK As Boolean
        Z = CboHusEdu.Items(0)
        ABI = CboHusEdu.Items(2)

        For x = 0 To CboHusEdu.Items.Count
            Try
                ABJ = CboHusEdu.SelectedItem.ToString()

                K = (String.Equals(ABJ, Z))
                ABK = (String.Equals(ABJ, ABI))
                If (K = True) Then
```

```
                    ' MessageBox.Show("    =output of selected index when K =" & K & " = " &
"String of Equal results /CboHusEdu.SelectedIndex.ToString/ =" &
CboHusEdu.SelectedIndex.ToString(0) & "  and " & CboHusEdu.SelectedItem.ToString())
                    ABK = K
                    Exit For
                ElseIf (ABK = True) Then
                    ' MessageBox.Show("    =output of selected index when U =" & ABK & " = "
& "String of Equal results /CboHusEdu.SelectedIndex.ToString/ =" &
CboHusEdu.SelectedIndex.ToString(0) & "  and " & CboHusEdu.SelectedItem.ToString())
                    K = ABK
                    Exit For
                End If
            Catch ex As Exception

            End Try
        Next

        Dim ABG As String
        Dim ABH As Boolean

        AC = CboANCVisit.Items(0)
        ABG = CboANCVisit.Items(5)

        For x = 0 To CboANCVisit.Items.Count
            Try
                AD = CboANCVisit.SelectedItem.ToString()

                AB = (String.Equals(AD, AC))
                ABH = (String.Equals(AD, ABG))

                If (AB = True) Then
                    ' MessageBox.Show("    =output of selected index when AB =" & AB & " = "
& "String of Equal results /CboANCVisit.SelectedIndex.ToString/ =" &
CboANCVisit.SelectedIndex.ToString(0) & "  and  " & CboANCVisit.SelectedItem.ToString())
                    ABH = AB
                    Exit For
                ElseIf (ABH = True) Then
                    ' MessageBox.Show("    =output of selected index when ABH =" & ABH & " =
" & "String of Equal results /CboANCVisit.SelectedIndex.ToString/ =" &
CboANCVisit.SelectedIndex.ToString(0) & "  and  " & CboANCVisit.SelectedItem.ToString())
                    AB = ABH
                    Exit For
                End If
            Catch ex As Exception
            End Try

        Next

        AL = CboSizeChild.Items(3)
        AO = CboSizeChild.Items(4)

        For x = 0 To CboSizeChild.Items.Count
            Try
                AM = CboSizeChild.SelectedItem.ToString()

                AK = (String.Equals(AM, AL))
```

```
                D = (String.Equals(AM, AO))

                If (AK = True) Then
                    ' MessageBox.Show("    =output of selected index when AK =" & AK & " = "
& "String of Equal results /CboSizeChild.SelectedIndex.ToString/ =" &
CboSizeChild.SelectedIndex.ToString(0) & "  and " & CboSizeChild.SelectedItem.ToString())
                    D = AK
                    Exit For
                ElseIf (D = True) Then
                    ' MessageBox.Show("    =output of selected index when D =" & D & " = " &
"String of Equal results /CboSizeChild.SelectedIndex.ToString/ =" &
CboSizeChild.SelectedIndex.ToString(0) & "  and " & CboSizeChild.SelectedItem.ToString())
                    AK = D
                    Exit For
                End If
            Catch ex As Exception
            End Try
        Next

        AY = CboThingOnCord.Items(0)
        For x = 0 To CboThingOnCord.Items.Count
            Try
                AZ = CboThingOnCord.SelectedItem.ToString()

                AX = (String.Equals(AZ, AY))
                If (AX = True) Then
                    ' MessageBox.Show("    =output of selected index when AX =" & AX & " = "
& "String of Equal results /CboThingOnCord.SelectedIndex.ToString/ =" &
CboThingOnCord.SelectedIndex.ToString(0) & "  and " & CboThingOnCord.SelectedItem.ToString())
                    Exit For
                End If
            Catch ex As Exception
            End Try
        Next

        If ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK = True
And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP =
True And ABF = True And A = True And ABB = True And ABC = True And C = True And F = True) =
True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = True And ABC = True And C = True And F = False)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = True And ABC = True And C = False And F = True)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
```

```
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = True And ABC = False And C = True And F = True)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = False And ABC = True And C = True And F = True)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = False And ABB = True And ABC = True And C = True And F = True)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = False And A = True And ABB = True And ABC = True And C = True And F = True)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = False And R = False And ABO = True And
ABP = True And ABF = True And A = True And ABB = True And ABC = True And C = True And F =
True) = True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = False And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = True And ABC = True And C = True And F = True) =
True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = False And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = True And ABC = True And C = True And F = True) =
True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
```

```
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = False And ABH = False And ABK
= True And K = True And U = True And W = True And B = True And R = True And ABO = True And
ABP = True And ABF = True And A = True And ABB = True And ABC = True And C = True And F =
True) = True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = False And D = False And AB = True And ABH = True And ABK
= True And K = True And U = True And W = True And B = True And R = True And ABO = True And
ABP = True And ABF = True And A = True And ABB = True And ABC = True And C = True And F =
True) = True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
False And K = False And U = True And W = True And B = True And R = True And ABO = True And
ABP = True And ABF = True And A = True And ABB = True And ABC = True And C = True And F =
True) = True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub

        ElseIf ((AX = False And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = True And ABP
= True And ABF = True And A = True And ABB = True And ABC = True And C = True And F = True) =
True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
False And K = False And U = True And W = True And B = True And R = True And ABO = True And
ABP = True And ABF = True And A = True And ABB = True And ABC = True And C = True And F =
True) = True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        ElseIf ((AX = True And AK = True And D = True And AB = True And ABH = True And ABK =
True And K = True And U = True And W = True And B = True And R = True And ABO = False And ABP
= False And ABF = True And A = True And ABB = True And ABC = True And C = True And F = True)
= True) Then
            MessageBox.Show("The likely of the under five children will be Dead")
            MessageBox.Show("Because of this chaild may have many complication so please
refere this mother for Hospital or higher clinc for further examination")
            Exit Sub
        Else
            MessageBox.Show("This mother can continuing her follow up at this health facility
with great care of practionars")
```

```
        End If
    End Sub

    Private Sub btnClose_Click(sender As System.Object, e As System.EventArgs) Handles
btnClose.Click
        Close()
    End Sub
    Private Sub Mother_FName_TextChanged(sender As System.Object, e As System.EventArgs)
Handles Mother_FName.TextChanged
        btnDelet.Enabled = True
        btnClose.Enabled = True
        btnNext.Enabled = True
        btnPrevious.Enabled = True
        btnSave.Enabled = True
        btnTestResult.Enabled = True
    End Sub

    Private Sub CboPlaceRes_SelectedIndexChanged(sender As System.Object, e As
System.EventArgs) Handles CboPlaceRes.SelectedIndexChanged

    End Sub

    Private Sub CboWmOccup_SelectedIndexChanged(sender As System.Object, e As
System.EventArgs) Handles CboWmOccup.SelectedIndexChanged

    End Sub

End Class
```