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

PREDICTING THE STATUS OF ANAEMIA IN WOMEN AGED 15-49 BY APPLYING DATA MINING TECHNIQUES USING THE 2011 ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY (EDHS) DATASETS

By

GETACHEW SETOTAW

JUNE 2013

ADDIS ABABA, ETHIOPIA

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

PREDICTING THE STATUS OF ANAEMIA IN WOMEN AGED 15-49 BY APPLYING DATA MINING TECHNIQUES USING THE 2011 ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY (EDHS) DATASETS

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

By

GETACHEW SETOTAW

JUNE 2013

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

PREDICTING THE STATUS OF ANAEMIA IN WOMEN AGED 15-49 BY APPLYING DATA MINING TECHNIQUES USING THE 2011 ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY (EDHS) DATASETS

By

GETACHEW SETOTAW

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chair person	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

Declaration

I declare that this thesis is my original work and that it has not been presented for a degree in this or any other university. Where other sources of information have been used, they have been acknowledged.

Name: _____

Signature: _____

Date: _____

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

Name: _____ **Name:** _____

Signature: _____ **Signature:** _____

Date: _____ **Date:** _____

ACKNOWLEDGEMENTS

I am sincerely and heartily grateful to my advisors, Dr Martha Yifiru and Professor Ahmed Ali, for the guidance, supports, and direction they showed me throughout my thesis writing. This thesis would not have been completed successfully without their continuous and invaluable helps.

I am also truly indebted and thankful to the Bahir Dar University for the generous supports through granting full Admission and Scholarships for my master's studies. I would like to show my earnest thankfulness to the Addis Ababa University for sponsoring this thesis research.

My special heartfelt thanks go to my wife Fikerte Getnet and to my daughters Bethelhem Getachew and Rodas Getachew, for their understanding and provision of unrestricted support and encouragement through both the ups and downs of my time during my study. My appreciation and thankfulness also go to my parents, brothers, and sisters, who morally and financially contributed to the completion of my study.

To all others who helped me in completing this study, I am equally grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

ANC	Antenatal Care
ANOVA	One-way Analysis of Variance
ARFF	Attribute Relation File Format
ART	Anti Retroviral Treatment
AUC	Area Under the Curve
BMI	Body-Mass Index
BRHP	Butajira Rural Health Programme
CART	Classification and Regression Tree
CBC	Complete Blood Cells
CHAID	Chi-squared Automation Interaction Detection
CLI	Command Line Interface
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CSA	Central Statistical Agency
CSV	Comma Separated Value
DALYs	Disability-Adjusted Life Years
DHS	Demographic and Health Survey
DM	Data mining
DT	Decision Tree
EAs	Enumeration Areas
EDHS	Ethiopia Demographic and Health Survey
FN	False Negative
FP	False Positive
g/dl	gram per deci litre
g/l	gram per litre
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
HSDP	Health Sector Development Program
KDD	Knowledge Discovery in Databases
MANOVA	Many-way Analysis of Variance
MIS	Management Information Science
NB	Naïve Bayes

NEFS	National Family and Fertility Survey
OLAP	On-Line Analytical Processing
RBC	Red Blood Cell
RDBMS	Relational Database Management System
ROC	Receiver Operating Characteristics
SAS	Statistical Analysis System
SEMMA	Sample Explore Modify Model Assess
SMO	Sequential Minimal Optimization
SNNP	Southern Nations, Nationalities and Peoples
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSRACT

Background: Anaemia is recognized as a major public health problem throughout the world. Globally, more than one-third of women and two-fifths of young children in the world are affected by anaemia. In Ethiopia, based on the 2005 and 2011 EDHS, 27% and 17% of women aged 15-49 are anaemic, respectively. Indeed, as per the WHOs classification on anaemia level, the prevalence of anaemia in Ethiopia is distinguished as moderate public health problem. Previous studies conducted using conventional statistical methods witnessed the public health problem of anaemia in Ethiopia.

Objective: The purpose of this research is to predict the status of anaemia in women aged 15-49 by applying data mining techniques using the 2011 EDHS dataset.

Methodology: The Knowledge Discovery in Database (KDD) which is applied to academic research was adopted to extract useful information from the dataset containing 6697 records. J48 Decision Tree and PART Rule induction were employed to build the predictive model. And also WEKA open software was used as a data mining tool to implement the experiments.

Results: The findings of this study revealed that all the models built using J48 decision tree and PART rule induction algorithms with all attributes have high classification accuracy and are generally comparable in predicting any anaemic cases. However, comparison that is based on the detailed performance measures suggests that the PART Unpruned model with all attributes perform better in predicting any anaemic cases with an accuracy of 95.2%.

Conclusion: The study showed that applying data mining techniques on the secondary data of EDHS to build a model that predicts any anaemic cases is possible. Thus, the outcome of this study helps health care planners and policy makers to design a proper and suitable preventive and control programs to combat anaemia.

Keywords: EDHS Dataset, Anaemia, Data Mining, Predictive Model

CHAPTER ONE

INTRODUCTION

1.1 Background

Anaemia is one of the most frequently observed nutritional deficiency diseases in the world today (1). It is especially prevalent in women of reproductive age, particularly during pregnancy when it is often a contributing cause of maternal health problem (2).

Anaemia refers to a condition characterized by a reduction in the red blood cell count or in the concentration of haemoglobin (3). Commonly, anaemia is the final outcome of nutritional deficiency of low iron intake, folate, and vitamin B_{12}, and some other nutrients (3). Malaria, sickle cell diseases, hemorrage, infection, parasite infestation (hookworm), chronic diseases are also considered as other significant causes of anaemia (1,3). In Africa, for instance, there are five common causes of anaemia, mainly, iron deficiency, folate deficiency, malaria, sickles cell disease, and HIV/AIDS (4).

Nutritional deficiency, due to a lack of bio-available dietary iron, accounts for the majority of anaemia in the world (5). The contribution of other causes of anaemia depends on many factors, including level of economic development, climate, condition of health care, and existence of anaemia control and prevention programs (3,5).

Globally, anaemia is considered as a worldwide public health problem (6), and children and women are the most affected groups (7). Anaemia in children has been associated with impaired cognitive performance, motor and language development, and scholastic achievement (2). Anaemia in women in turn reduces their work productivity and places them at risk of poor pregnancy outcome, including increased risk of maternal mortality, prenatal mortality, premature births, and low-birth weight (1).

Recognizing anaemia as a major public health problem worldwide, most of the member states of WHO conducted a population based survey such as the Demographic and Health Survey (DHS) to get an estimate of anaemia prevalence among vulnerable groups. The resulting data could provide important background information for public health policy decisions that are essential for the development of national and community-based anaemia prevention programs (3).

In Ethiopia, so far three Demographic and Health Surveys (EDHS) have been conducted at five-year intervals since 2000 with primary objectives of providing quality information for planning, policy formulation, monitoring and evaluation of population, and health programs in the country. The 2005 and 2011 EDHS were the second and third surveys respectively which presented results on HIV and anaemia prevalence's (8).

According to 2005 EDHS (9), the primary purpose of conducting the DHS in Ethiopia was to furnish policy makers, planners, researchers, and program managers with detailed information of the population and health situation, covering topics on family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, malaria, women's empowerment, and knowledge of HIV/AIDS. In addition to this information, the 2005 and 2011 EDHS include population based estimates of HIV and anaemia prevalence in the country which makes them richer than the 2000 EDHS (8, 9).

According to Kantardzic (11), in the health sector, large amounts of data are becoming available due to the widespread use of information technologies. These an extensive amount of data then require specialized tools and methods that can be used to discover hidden knowledge which is useful in decision making and problem solving. Nowadays, among such specialized tools applicable in the health care sector, data mining technique is the required one (11). Data mining is an emerging technology that has made a revolutionary change in the information world (12). It serves as a means of searching previously unknown, actionable information from large databases (13).

Data mining, in the health care sector, is set to play an important role in tackling the data overload (14). In this regard, Daniel (13) states that benefits of data mining include improving health care quality, reduced operating costs, and better insight into health data. Given the benefits acquired in applying data mining techniques in the health care sector, it is proper to assess the relevance and potential advantage of such emerging technology in the Ethiopian health sector.

Thus, since there is still a moderate prevalence of anaemia in Ethiopia, in this study the researcher is motivated to explore the potential applicability of data mining techniques to

predict prevalence of anaemia in women of reproductive age in Ethiopia by using the Ethiopia Demographic and Health Survey of 2011 datasets.

1.2 Statement of the problem

Anaemia is recognized as a major public health problem throughout the world (3). And occurs at all stages of the life cycle, but is more prevalent in pregnant women and young children (2). Globally, the most significant contributor to the onset of anaemia is iron deficiency (5). In relation to this, the WHO/World Bank supported analysis of the Global Burden of Disease ranked iron deficiency anaemia as the third leading cause of loss of disability-adjusted life years (DALYs) for females aged 15-44 across the globe (3).

According to the epidemiological data collected from multiple countries of DHS by the World Health Organization (WHO), more than one-third of women and two-fifths of young children in the world are affected by anaemia (1). In developing countries, about half of women and young children are anaemic (1). The highest overall rates of anaemia are reported in South Asia and in certain regions of Africa (6).

In Ethiopia, based on the 2005 and 2011 EDHS surveys, 27% and 17% of women aged 15-49 were anaemic respectively. Indeed, as per the WHO (2) classification on anaemia level, the prevalence of anaemia in women of Ethiopia is distinguished as moderate public health problem (8,9). In both surveys, anaemia status was reported by haemoglobin level ranging from any anaemia: mild, moderate, and severe anaemia. Although the data on anaemia prevalence is used as good starting point to track progress towards the target of reducing anaemia prevalence, anaemia still remains a problem in Ethiopia since those anaemia prevalence surveys had not properly addressed the major determinants of anaemia and were limited to solely measuring haemoglobin level to present the results on anaemia prevalence (8-10).

The determination of factors that influence the occurrence of anaemia in women of reproductive age is fundamental for the implementation of control measures (9). In view of this, to determine the prevalence of anaemia among women of reproductive age, and to explore some factors commonly associated with anaemia, several studies were conducted over the years that indicate the public health problem of anaemia in women of Ethiopia (15). Haider (16) reported that 30.4% of Ethiopian women were anaemic. A study by Haider and

Pobock (17) presented the overall prevalence rate of iron-deficiency anaemia among women were 18%. Samson and Fikre (15) also reported that 27.4% of Ethiopian women of reproductive age were anaemic. Most of these previous studies used factors such as socio-economic, demographic, nutritional, and reproductive, apart from health related variables, to identify the determinant risk factors of anaemia in women of reproductive age (15).

However, in addition to being small proportion of the datasets, these studies were conducted using conventional statistical methods which usually have limited capacity to discover new and unanticipated patterns that are hidden in data. In such situations, new information technologies like data mining may be helpful in discovering hidden data patterns.

Thus, besides the existence of a public health problem of anaemia in women of Ethiopia, the absence of research using data mining techniques on the problem domain of previous studies motivate this study to investigate the potential applicability of data mining techniques to predict the status of anaemia in women of reproductive age based on the secondary data of the Ethiopia Demographic and Health Survey (EDHS) 2011 datasets.

1.3 Objective of the study

1.3.1 General objective

The general objective of this research is to predict the status of anaemia in women aged 15-49 by applying data mining techniques using the 2011 EDHS survey datasets.

1.3.2 Specific objectives

To achieve the general objective of this research, the following specific objectives are envisaged:

- To build predictive models in order to classify women of reproductive age to any anaemic/not anaemic levels,
- To compare and evaluate the performance of the resultant predictive data mining models, and select the one which performs the best,
- To discover the potential set of attributes which can reasonably predict any anaemic women,

1.4 Research Methodology

1.4.1 Study setting

Ethiopia is situated in the horn of Africa. It is a country with great climatic, geographic and cultural diversity. According to the 2007 census, it has a population of 73.8 million of which only 16 percent of the population lives in urban areas (18). The majority of the population (84%) lives in the rural areas. More than one fifth (23.3%) of the population are women in the reproductive age (18). The country is characterized by an average annual population growth rate of 2.6 percent and a total fertility rate of 4.8 children per women (8, 18).

The 2011 EDHS survey covers both urban and rural areas of Ethiopia. A total of 11 geographic/administrative regions, that is, nine regional states and two city administrations, namely: Afar, Amhara, Gambella, Harrari, Oromia, Somali, Benshangul-Gumuz, Southern Nations, Nationalities and Peoples (SNNP), Tigray, and Addis Ababa, and Dire Dawa were included in the study(8).

1.4.2 Data collection

This research was generally based on secondary data extracted from the Ethiopia Demographic and Health Survey (EDHS) of 2011 datasets which was conducted by the Central Statistical Agency (CSA) under the worldwide MEASURE DHS project. This survey is a valid dataset and is available to researchers through the Ethiopian Central Statistical Agency (CSA) on formal request and the MEASURE DHS project website at http://www.measuredhs.com. The researcher obtained such EDHS 2011 datasets from the Ethiopian Central Statistical Agency (CSA) after submitting a formal letter written by the School of Information Science.

In this study, three files such as Household, Individual, and Couples' stored in the 2011 EDHS dataset were used as a source of the data. All these files were available in SPSS format contained a total of 45, 539 records (rows) and 920 attributes (columns).

1.4.3 Study design

This study employed a knowledge Discovery in Database (KDD) method to build predictive models using data mining techniques by applying a set of classifiers algorithms on the EDHS

2011 datasets. This method was selected for this specific study because of various reasons. First, the KDD methodology was best suited for academic researches. Second, using KDD methodology in such research reduces the skill required for the knowledge discovery. Third, the KDD methodology involves an iterative feature of the steps during the study.

The current research followed the five steps of the knowledge Discovery in Database (KDD) methodology, mainly, data selection, data pre-processing, transformation, model building, and performance evaluation of models which is based on data mining tools. These steps were iterative, with many decisions made by the user. Note that the process is iterative at each step, meaning that moving back to previous steps may be required.

To build supervised predictive models, the study exploited J48 Decision Tree and PART Rule Induction algorithms. For the purpose of this work, the researcher used the tool WEKA open software for the analysis. To compare and evaluate the performance of the models, the standard performance metrics such as Correctness Accuracy, TP Rate, TN Rate, Precision, Recall, ROC Curve, Speed, and Confusion matrix were used.

1.5 Scope of the Study

The scope of this study was limited to developing a predictive model that is capable of predicting anaemic cases in women of reproductive age using data mining tools and techniques, specifically classification techniques. Although the tool selected for this study, which is WEKA software, provides many classification and prediction techniques, this study was restricted to applying only J48 Decision Tree and PART Rule Induction algorithms.

The scope of this research was also restricted to utilizing the secondary data of the 2011 Ethiopian Demographic and Health Survey (EDHS) dataset. Thus, only the raw data of women aged 15-49 years which was stored in the 2011 EDHS was employed for the study.

1.6 Significance of the study

The significance of the study is as follows:

- The study has a paramount importance for public health planners to plan and implement preventive actions focused on maternal health strategies so as to mitigate the risk of anaemia attributed by avoidable recommended factors.

- It will assist health care planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at preventing and treating anaemia in the country.
- It will be useful for further consultancy for research.

1.7 Dissemination of Results

The result of this research will be disseminated through the following ways:

- Preparation of the Thesis
- Presentation at the thesis defence program
- Publishing in local and international journals
- Presentation in different conferences/workshops/symposium
- Putting the hardcopy and the softcopy in the Health Informatics library and other libraries so that interested readers can get access to the research output to be used for decision support, take action or to use it as a base for further research in the area.

1.8 Structure of the Thesis

The thesis is organized into six chapters.

The first chapter is an introductory chapter which gives a brief background about the study. It also describes the problem that constitutes the basis of the paper, objectives of the study, the research methodology, significance and scope of the study, and finally dissemination of the results.

The second chapter deals with reviewing literatures conceptually relevant and related to this study. Chapter three is concerned with understanding the application domain and the dataset selected for the study.

The fourth chapter of this thesis is devoted to describing the overall research design of the study. In this chapter, the five consecutive steps of the KDD including data preparation are presented in detail.

The fifth chapter focuses on modelling and analysis of the results. It provides discussion of research findings in detail along with their implications

Finally, in chapter six, the conclusion and recommendations are addressed. In this chapter, the thesis concludes the study in four parts: summary of the thesis, answers to the research objectives, strengths and limitation of the study, and finally recommendations.

CHAPTER TWO

LITERATURE REVIEW

It is essential for any research undertaking to review previous studies in the area of investigation and sum up the trends of research practices and the direction of the findings there from. In this chapter, literatures conceptually relevant and related to this study have been reviewed.

2.1 Review of Literatures

2.1.1 The Foundations of Data Mining

Data mining can be viewed as a result of the natural evolution of information technology (14). This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery (19).

In the business community as noted by Berson et al. (20), data mining was ready for application with the support of three technologies, namely, massive data collection, powerful multiprocessor computers, and data mining algorithms that are now sufficiently mature. Among these three technologies, data mining algorithms embody techniques that have only recently been implemented as mature, reliable, understandable tools consistently outperforming older statistical methods (20).

According to Thearling (19), in the steps of evolution of data mining (Table 2.1) from business data to business information, each new step is built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining.

Table 2.1: Steps in the Evolution of Data Mining.

Evolutionary Step	Enabling Technologies	Characteristics
Data Collection (1960s)	Computers, tapes, disks	Retrospective, static data delivery
Data Access (1980s)	Relational databases (RDBMS), Structured Query Language (SQL)	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (2000+)	Advanced algorithms, multiprocessor computers, massive databases	Prospective, proactive information delivery

2.1.2 Overview of Data Mining

The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data (14). Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing) (12). Among such names, the term "data mining" is primarily used by statisticians, database researchers, and the MIS and business communities (12).

Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to: numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, and neural networks and genetic algorithms (11). Consequently, data mining consists of more than collecting and managing data. It includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data (21). Such parameters as noted by Two Crows Corporation (21) include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper); sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers); classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases); clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences); and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).

While many data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is commonplace to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships. In this regard, Daniel (13) wrote, as compared to other data analysis applications, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. In contrast, many simpler analytical tools or statistical analysis software utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis (13).

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source (21).

In addition to these improved data management tools, the increased availability of information and the decreasing costs of storing it have also played a role. Over the past several years there has been a rapid increase in the volume of information collected and

stored, with some observers suggesting that the quantity of the world's data approximately doubles every year (21). At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease (21).

Reflecting such conceptualization of data mining, various definitions of, and synonyms for, data mining have emerged in recent years. Translating data mining word by word means the mining or digging in data with the purpose of finding useful information or respectively knowledge. Coming to the more abstract and very well known definition of Moxon (22), data mining is defined as "The non-trivial extraction of implicit, previously unknown, and potentially useful information from data".

Groth (23) mentions another interesting aspect of data mining. He describes it as "a knowledge discovery process of extracting previously unknown, actionable information from very large dataset". In the words of Witten and Frank (24): "Data mining is the process of meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques". As pointed out by Witten and Frank (24), the idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data.

Apart from the definitions of data mining, many researchers and practitioners treat data mining as a synonym for another popularly used term Knowledge Discovery in Databases, or KDD, while others view data mining as simply an essential tool in the process of knowledge discovery process (12). In relation to this, some observers consider data mining to be just one step in a larger process known as Knowledge Discovery in Databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation (14).

2.1.3 Uses of Data Mining

Data mining tools can predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analysis offered by data

mining move far beyond the analysis of past events provided by retrospective tools typically offered by decision support systems (19).

Data mining tools can give answers to business questions that traditionally were time consuming to resolve (19). Today data mining is primarily used by companies such as banking, insurance, medicine, retailing, and health care to reduce costs, enhance research, and increase sales. Such companies, however, use data mining for different purposes (12). According to David (12), a few areas in which information industries use data mining to achieve a strategic benefit are:

- the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely.
- the medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine.
- Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases.
- Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor.

In healthcare organizations, there are several factors that have motivated the use of data mining applications (25). Among other factors, as identified by Milley (25) is that:

- the existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders.
- the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data.

- the realization that data mining can generate information that is very useful to all parties involved in the healthcare industry. For example, data mining applications can help healthcare insurers detect fraud and abuse, and healthcare providers can gain assistance in making decisions, for example, in customer relationship management. Data mining applications also can benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices.

Recently, data mining has been increasingly cited as an important tool for health care organizations. Accordingly, Eapen (26) observed that "data mining techniques are being increasingly implemented in healthcare sectors in order to improve work efficiency and enhance quality of decision making process". In addition, the use of data mining in healthcare, as suggested again by Eapen, (26), enables comprehensive management of medical knowledge and its secure exchange between recipients and providers of healthcare services; the elimination of manual tasks of data extraction from charts or filling of specialized questionnaires; extraction of data directly from electronic records; transfer on secure electronic system of medical records that will save lives and reduce the cost of health care, early detection of infectious diseases with advanced collection of data etc.

2.1.4 Data Mining and Statistics

The disciplines of statistics and data mining both aim to discover structure in data. So much do their aims overlap, that some people regard data mining as a subset of statistics. But that is not a realistic assessment as data mining also makes use of ideas, tools, and methods from other areas (27). Friedman (27) further described that data mining is used to discover patterns and relationships in data, with an emphasis on large observational databases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated exploratory data analysis of (usually) large complex data sets.

Statistical procedures do, however, play a major role in data mining, particularly in the processes of developing and assessing models. Most of the learning algorithms use statistical

tests when constructing rules or trees and also for correcting models that are over-fitted. Statistical tests are also used to validate machine learning models and to evaluate machine learning algorithms (28).

The statistical analysis procedures provided by current data mining packages nearly always include: Decision tree induction (C4.5, CART, CHAID); Rule induction (CN2, Recon, etc.); Neural networks; Bayesian belief networks ("graphical models"); Nearest neighbors ("case based reasoning"); Clustering methods ("data segmentation"); Association rules ("market basket analysis"); Feature extraction; and Visualization (27).

Almost none of these data mining packages offer: Hypothesis testing; Experimental design; Response surface modelling; ANOVA, MANOVA, etc.; Linear regression; Discriminant analysis; Logistic regression; Canonical correlation; Principal components; and Factor analysis (27). These latter procedures are of course the mainstay of the standard statistical packages. Thus, nearly all of the methodology currently being marketed (and used) for data mining has been developed and promoted in fields other than statistics (27, 28).

2.1.5 Methodologies for Data Mining

As data mining is coming of age, several methodologies have been developed, each with their own perspective (11). Broadly used methodologies in data mining as described by Daniel (13) are: KDD (Knowledge Discovery in Data base), CRISP-DM (Cross-Industry Standard Process for Data Mining), and SEMMA (Sample Explore Modify Model Assess) process. Despite that KDD, CRISP-DM, and SEMMA are usually referred as methodologies, they are also referred as processes, in the sense that they consist of a particular course of action intended to achieve a result (13). Here all of them are reviewed as follows:

2.1.5.1 KDD Methodology

The term knowledge discovery in databases or KDD, for short, was coined in 1989 to refer to the broad process of finding knowledge in data, and to emphasize the "high-level" application of particular data mining techniques (21). Kantardzic (11) defined the KDD process as: "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Unlike others methodologies, the KDD process is

categorized as an academic model which is considered to be more applicable for academic research works.

The approach to gain knowledge out of a set of data was separated by Fayyad (29) into individual steps. The individuality results out of different tools we use, and different outcomes that are needed. According to Fayyad (29) there are five KDD steps: Selection, Pre-processing, Transformation, Data Mining, and Interpretation/Evaluation. These five steps are passed through iteratively. Every step can be seen as a work-through phase. Such a phase requires the supervision of a user and can lead to multiple results. The best of these results is used for the next iteration, the others should be documented. An outline of the steps that are in Figure 2.1 will be adequate for understanding the concepts required for the KDD process.

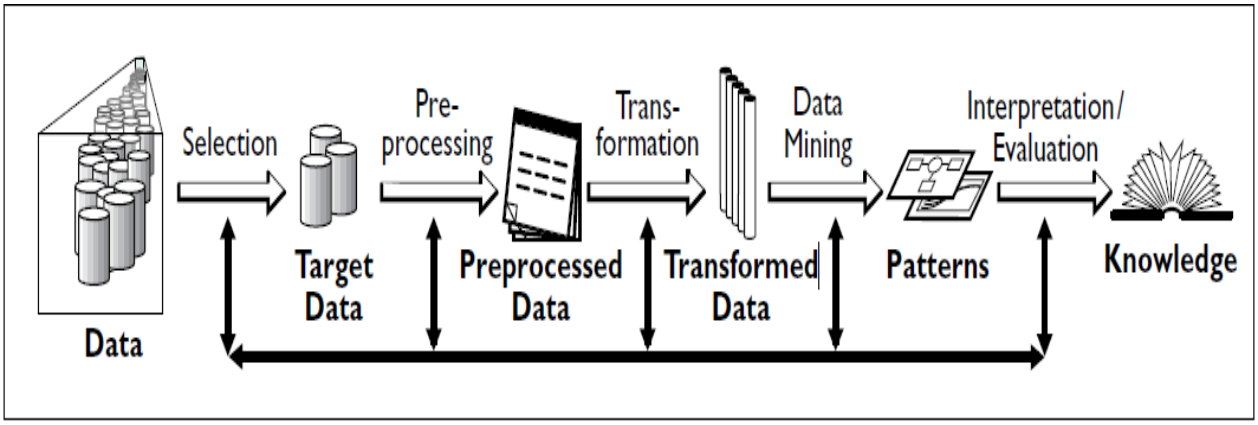


Figure 2.1: The Knowledge Discovery in Database (KDD) Process

Next, the steps will be briefly described.

Step 1: Selection

In the selection-step the significant data gets selected or created. Henceforward the KDD process is maintained on the gathered target data. Only relevant information is selected, and also metadata or data that represents background knowledge in a case where we have millions of data points.

Step 2: Pre-processing

A good result after applying data mining depends on an appropriate data preparation in the beginning. Important elements of the provided data have to be detected and filtered out. These kinds of things are settled in the pre-processing phase. To detect knowledge the effective main task is not only to apply data mining but also to pre-process the data properly and tools. The less noise contained in data the higher is the efficiency of data mining. As noted by Fayyad (29), elements of the pre-processing include the cleaning of wrong data, the treatment of missing values, and the creation of new attributes.

Step 3: Transformation

Transformation of data in this step can be defined as decreasing the dimensionality of the data that is sent for data mining. With the reduction of dimensionality we increase the efficiency of the data-mining step with respect to the accuracy and time utilization. The transformation phase of the data may result in a number of different data formats, since variable data mining tools may require variable formats. The data can be reduced either manually or automatically. The reduction can be made via lossless aggregation or a loss full selection of only the most important elements. A representative selection can be used to draw conclusions to the entire data.

Step 4: Data Mining

The data mining step is the major step in KDD. In the data mining phase, the data mining task is approached. This is when the cleaned and pre-processed data is sent into the intelligent algorithms for classification, clustering, similarity search within the data, and so on. Here we choose the algorithms that are suitable for discovering patterns in the data. Some of the algorithms provide better accuracy in terms of knowledge discovery than others. Thus selecting the right algorithms can be crucial at this point.

Step 5: Interpretation/Evaluation

This stage consists on the interpretation and evaluation of the mined patterns. This indicates interpreting the discovered patterns and possibly returning to any of the previous steps, as

well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users. The evalution process is also carried out to identify interesting patterns representing knowledge based on some interestingness measures.

2.1.5.2 CRISP-DM Methodology

CRISP-DM stands for CRoss-Industry Standard Process for Data Mining. It evolved to become the *de facto* industry standard. The CRISP-DM process was developed by the effort of a data mining companies to provide a generic process model that can be specialized according to the needs of any particular company or industry (30). It consists in a cycle that comprises six stages as illustrated in Figure 2.2:

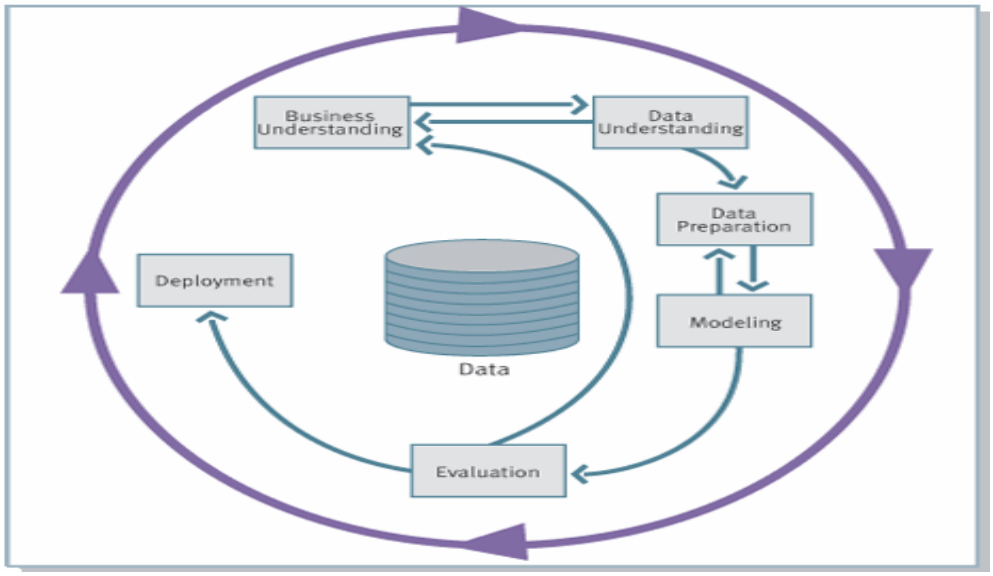


Figure 2.2: Phases of the CRISP-DM process Model

The following is a brief description of each of the phases in CRISP-DM cycle.

Phase 1: Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining 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 to become 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 (data that will be fed into the modeling tool(s) from the initial raw data). Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Phase 4: Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements regarding the form of data. Therefore, stepping back to the data preparation phase is often needed.

Phase 5: Evaluation

At this stage in the project the model (or models) built appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been considered sufficiently. At the end of this phase, a decision on the use of the data mining results should be reached.

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 client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the client to understand up front what actions will need to be carried out to make use of the models (30).

2.1.5.3 The SEMMA Analysis Cycle

SEMMA, like CRISP-DM, grow as industrial standard and define a set of sequential steps that pretends to guide the implementation of data mining applications. SEMMA stands for the five steps of the analyses, namely, Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project as shown in Figure 2.3.

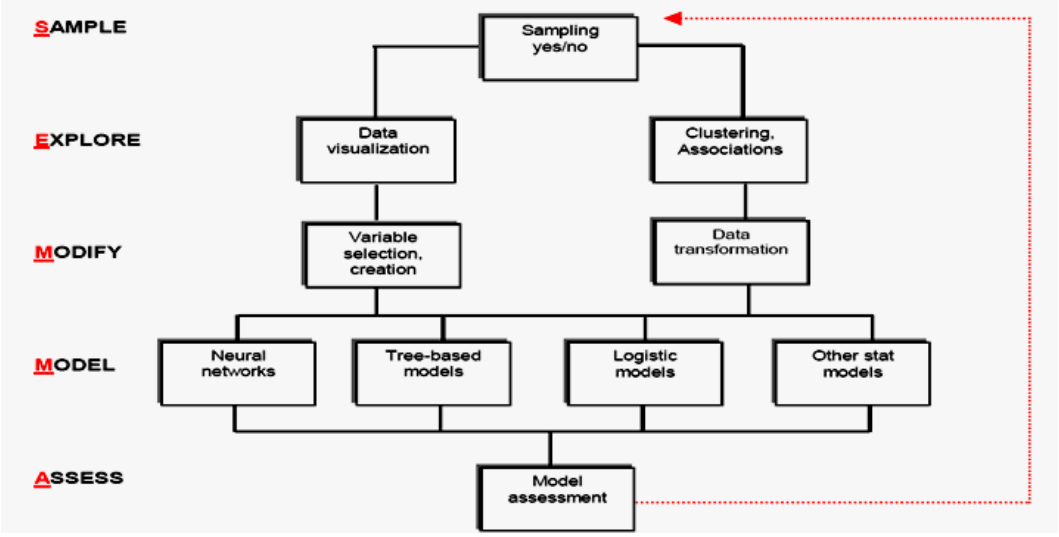


Figure 2.3: The SEMMA Analysis Cycle.

The following is a brief description of each of the steps in SEMMA process.

Step 1: Sample

This step deals with of sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.

Step 2: Explore

After sampling the data, the next step is to explore the data visually or numerically for trends or groupings. Exploration helps to refine the discovery process. Techniques such as factor analysis, correlation analysis and clustering are often used in the discovery process.

Step 3: Modify

This step deals with on the modification of the data by creating, selecting, and transforming the variables to focus on the model selection process, or to modify the data for clarity or consistence.

Step 4: Model

This step concerned with on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Step 5: Assess

The last step is to assess the model to determine how well it performs. A common means of assessing a model is to set aside a portion of the data during the sampling stage. If the model is valid, it should work for both the reserved sample and for the sample that was used to develop the model (31).

2.1.6 Data Mining Tasks

There are different kinds of data mining functionalities (tasks) that can be used to extract knowledge from large database. These could be classified into two main categories: “Descriptive” and “Predictive” which is considered as the high level goals of data mining (11,14).

On the descriptive end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. Clustering,

Summarization, and Visualization of databases are the main applications of descriptive data mining. The usefulness of this concept is that it enables one to generalise the data set from multiple levels of abstraction, which facilitates the examination of the general behaviour of the data, since it is impossible to deduce that from a large database.

On the other hand, in predictive end of the spectrum, the goal of data mining is to produce models that are capable of producing prediction results when applied to unseen cases. Classification and Regression are the most frequent types of tasks that are applied in predictive data mining. However, the relative importance of prediction and description for particular data-mining applications can vary considerably (14).

So, since this research project aims on predictive data mining, one of its types a classification task suitable for this study is discussed below.

2.1.6.1 Classification

Classification, as indicated before, is a task in predictive data mining. According to Han and Kamber (14), "classification is 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". The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Usually, in the classification process, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it (14).

In general, as described by Kantardzic (11), classification is the process of assigning a discrete label value (class) to an unlabeled record, and a classifier is a model (a result of classification) that predicts one attribute - class of a sample - when the other attributes are given. In doing so, samples are divided into predefined groups. For example, a simple classification might group customer billing records into two specific classes: those who pay their bills within thirty days and those who take longer than thirty days to pay. Different classification methodologies are applied today in almost every discipline where the task of classification, because of the large amount of data, requires automation of the process.

Examples of classification methods used as a part of data-mining applications include classifying trends in financial market and identifying objects in large image databases (11).

As Han and Kamber (14) wrote, classification is a two-step process: Model construction and model usage. The former step is concerned with the building of a classification model by describing a set of predetermined classes using training data set. The training data set a set of tuples where each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The learned model is represented in the form of classification rules, decision trees, or mathematical formula. The later step involves the use of a model (classifier) built to predict or classify unknown objects based on the patterns observed in the training set.

There are various classification methods. Popular classification techniques include the following (13,14):

- Decision Tree based Methods,
- Rule-based Methods (Rule Induction),
- Neural Networks ,
- Bayesian Networks,
- K-Nearest Neighbour,
- Support Vector Machines.

Although the choice of techniques suitable for classification tasks seems to be strongly dependent on the application, the data mining techniques that are applied in many real-world applications as a powerful solution to classification problems, among other methods, are decision trees and rule inductions (19). This section, therefore, discusses the concepts and principles of these techniques namely decision trees and rule inductions.

2.1.6.1.1 Decision Trees

Decision tree is well-known to be an effective classification task in several domains (19). It is a way of representing series of rules that lead to a class or value. Decision tree, as defined by Han and Kamber (14), "is a flow-chart-like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. In general, it is a knowledge representation structure consisting

of nodes where the top most nodes are the root node and branches organized in the form of a tree such that, every internal non-leaf node is labelled with values of the attributes (11).

Decision tree is one of the easier data structure to understand data mining, and also one of the most widely used and practical forms of machine learning and data mining. Decision tree algorithms are based on a divide-and-conquer approach to the classification problem. They work from the top down, seeking at each stage an attribute to split on that best classes; then recursively processing the sub-problems that result from the split. This strategy generates a decision tree, which can if necessary be converted into a set of classification rules—although if it is to produce effective rules, the conversion is not trivial (24).

A decision tree is necessarily a tree with an arbitrary degree that classifies instances. In decision trees, the leaf node represents the complete classification of a given instance of the attribute and the decision node specifies the test that is conducted to produce the leaf node. Thus with a decision tree, the sub tree that is created after any node is necessarily the outcome of the test that was conducted. From Figure 2.4 below, it is noted that the decision node is actually an attribute, which is characterized by the values present in it to describe a symptom or take a decision.

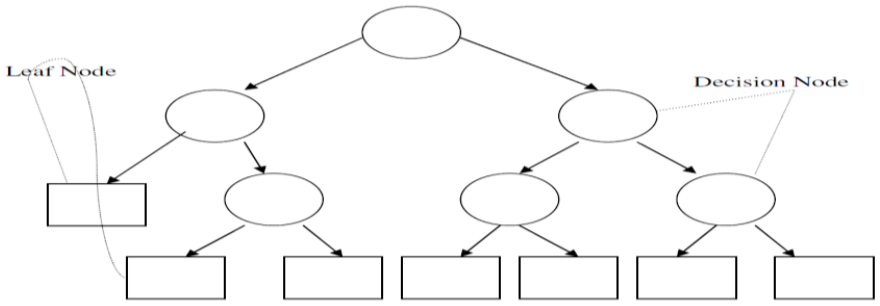


Figure 2.4: An Example of Decision Tree

A decision tree is a hierarchical structure with each node containing decision attribute and node branches corresponding to different attribute values of the decision node. The goal of building decision tree is to partition data with mixing classes down the tree until the leaf nodes contain pure class. A major issue in using decision tree is to find out how deep the tree should grow and when it should stop. Usually if all the attributes are different and lead to the

same outcome, the decision tree might not be the most effective in making decision and, at the same time, the size of the tree will be large (14).

A decision-tree model is built by analyzing training data and the model is used to classify unseen data. However, when decision trees are built, many of the branches may reflect noise or outliers in the training data. Considering the goal of research and improving classification accuracy of unseen data, tree pruning was attempted in order to identify and remove unnecessary branches that lead to over-fitting of the model. There are a number of algorithms that are based on decision trees. Some of the most common and effective types of algorithms based on decision trees are C4.5, FACT and Classification and Regression Tree (CART) (24).

2.1.6.1.2 Rule Induction

The second algorithm is rule induction method which has been found to be comparable in performance with decision tree (19). Rule induction is one of the major forms of data mining and is perhaps the most commonly used data mining techniques for classification tasks (19). It is also perhaps the form of data mining that most closely resembles the process that most people think about when they think about data mining, namely "mining" for gold through a vast database. The gold in this case would be a rule that is interesting - that tells something about the database that we didn't know and probably weren't able to explicitly articulate (aside from saying "show me things that are interesting") (32).

Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of data dining. Usually rules are expressions of the form:

***IF* (*attribute*−1, *value*−1) *and* (*attribute*−2, *value*−2) *and* · · ·**
***and* (*attribute*−*n*, *value*−*n*) *THEN* (*decision*, *value*).**

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain. Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labelled as attributes and a

decision (33). The representative rule induction method, which is available in WEKA includes: Prism, Ripper, PART, OneR.

2.1.6.1.3 Decision trees Vs. Rule inductions

In this section, the common difference and similarities existed between decision trees and rule inductions have been discussed.

Decision trees also produce rules but in a very different way than rule induction systems (19). The main difference between the rules that are produced by decision trees and rule induction systems as pointed out by Thearling (19) is that "Decision trees produce rules that are mutually exclusive and collectively exhaustive with respect to the training database while rule induction systems produce rules that are not mutually exclusive and might be collectively exhaustive".

Thearling (19) noted that the reason for this difference is the way in which the two algorithms operate, that is, rule induction seeks to go from the bottom up and collect all possible patterns that are interesting and then later use those patterns for some prediction target. Decisions trees, on the other hand, work from a prediction target downward in what is known as a "greedy" search.

Regarding their commonality, Thearling (19) again stated that "the one thing that decision trees and rule induction systems have in common is the fact that they both need to find ways to combine and simplify rules". In a decision tree this can be as simple as recognizing that if a lower split on a predictor is more constrained than a split on the same predictor further up in the tree that both don't need to be provided to the user but only the more restrictive one. While rules from rule induction systems are generally created by taking a simple high level rule and adding new constraints to it until the coverage gets so small as to not be meaningful.

2.1.7 Data Mining Tool

There are different data mining software's that can be used to build and test predictive models. Among those tools the WEKA which was selected for this study is discussed below.

2.1.7.1 WEKA Software

WEKA open software is one of the data mining tools that are used for data mining research purposes. WEKA was developed at the University of Waikato in New Zealand, and the name stands for *Waikato Environment for Knowledge Analysis*. A number of data mining methods are implemented in the WEKA software. Some are based on decision trees like the J48 decision tree, and some are rule-based like PART and decision tables, and others are based on probability and regression, like the Naïve Bayes algorithm (24).

The WEKA software is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems—and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset. WEKA is organized in packages that correspond to a directory hierarchy. It consists of four graphical user interfaces such as Explorer, Experimenter, Knowledge Flow and Simple Command-Line Interface (24) as shown in Figure 2.5. According to Witten and Frank (24), the easiest way to use WEKA is through a graphical user interface called the Explorer which gives access to all of its facilities using menu selection and form filling. Thus, most people choose the Explorer than others, at least initially.

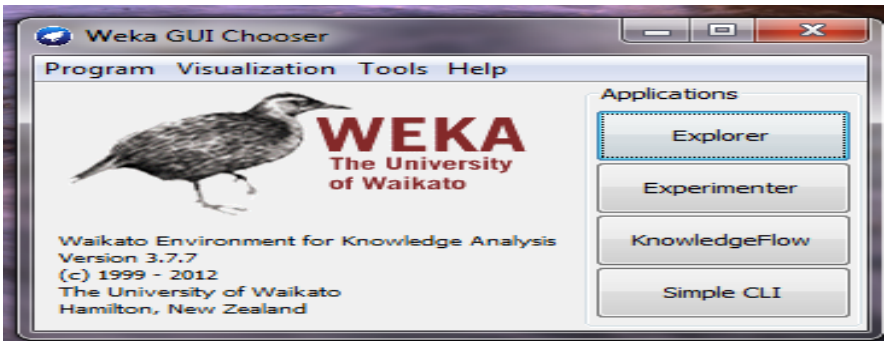


Figure 2.5: The WEKA Graphical User Interfaces

2.1.8 Model Evaluation

Evaluation is the most important step in any data mining task and the key to making real progress in data mining, because it gives us a clear view regarding the strength of each model individually, and enables us to compare different models based on common performance

measures (34). This step determines whether or not the models are working. According to Weiss and Zhang (34), a model performance evaluation should answer questions such as:

- How accurate is the model?
- How well does the model describe the observed data?
- How much confidence can be placed in the model's predictions?
- How comprehensible is the model?

Of course, the answer to these questions depends on the type of model that was built. Evaluation here refers to the technical merits of the model, rather than the measurement phase of the virtuous cycle (34).

Performance of models is analyzed based on a series of measures. The standard performance measures that were used for evaluation of include: confusion matrix such as correctness accuracy, sensitivity (TP Rate), specificity (TN Rate), precision, recall, F-measure, and ROC (Receiver Operating Characteristic). Each of these was used where appropriate in the analysis of the performances (34).

2.1.8.1 Confusion Matrix

In evaluating the performance of a model, a confusion matrix or correct classification matrix can be used. Confusion matrix focuses on the predictive capability of a model rather than how fast it takes to classify or build models, scalability, etc. (24). In the class case prediction, the result is often displayed as a two dimensional confusion matrix with a row and column for each class. Figure 2.6 shows the two dimensional confusion matrix which contains both predicted and actual classes.

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

Figure 2.6: Two dimensional Confusion Matrix

In the two-class case with classes yes and no, a single prediction has the four different possible outcomes shown in Figure 2.7. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Figure 2.7: Different outcomes of a two-class prediction.

Among the standard performance metrics, accuracy is the most widely-used metric to check the performance of the model. The correctness accuracy for a data mining classifier is defined as the degree of closeness of its prediction to the actual values, either true or false (34). The accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives instance by the total number of samples.

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.1)$$

The performance of the model that enables it to classify the positive cases correctly is sensitivity. Sensitivity is defined as its ability to correctly identify actual cases, i.e. for this study, it measures the proportion of any anaemic instances which are correctly identified as such. The performance of the model to classify the negative cases is specificity. Specificity is defined as its ability to correctly identify negative cases which are correctly identified as such. They are summarized in the following formulas (34):

$$Sensitivity = \frac{TP}{TP+FN} \quad (2.2)$$

$$Specificity = \frac{TN}{TN+FP}$$

Another metrics for performance evaluation of the classifier is precision which measures what percent of tuples that the classifier labelled as positive are actually positive. Precision is calculated by dividing correctly classified instances by the total number of correctly and incorrectly classified samples.

$$Precision = Positive\ Predictive\ Value = \frac{TP}{TP + FP} \qquad (2.3)$$

Recall is also another performance evaluation which measures what percent of positive tuples the classifier labelled as positive for both True and False classes. The formula is:

$$Recall = Sensitivity = TP\ Rate = \frac{TP}{TP + FN} \qquad (2.4)$$

The final metric used for performance evaluation of classifiers on confusion matrix is F-measure. The F- measure is the inverse relationship between precision & recall, and calculated as the harmonic mean of recall and precision. It is the point to conclude that the precision and recall of the model are significantly balanced (34).

$$\text{F-measure}(\text{F}) = \frac{2rp}{r + p} = \frac{2TP}{2TP + FN + FP} \qquad (2.5)$$

2.1.8.2 ROC Analysis

An alternative method that was used to evaluate classifier performance is the Receiver Operating Characteristic (ROC) analysis (14). ROC compares visually the relative performance among competing models. In ROC analysis, performance of each classifier represented as a point on the ROC curve. The ROC curve shows the trade-off between the true-positive (TP) rate (proportion of positive tuples that are correctly identified) and the false-positive (FP) rate (proportion of negative tuples that are incorrectly identified as positive) for a given model (14).

The ROC curve plots the true positive rate on the vertical axis and the false positive rate on the horizontal axis as shown in Figure 2.8.

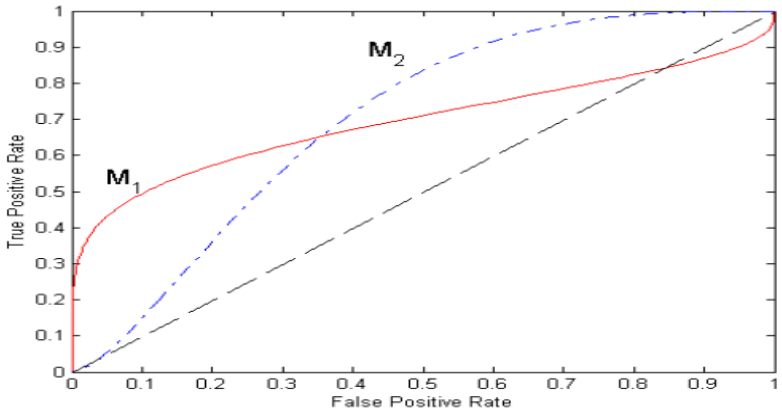


Figure 2.8: A Sample ROC curves

As a standard method for evaluating classifiers, the primary advantage of ROC curve is that they are used to evaluate the performance of a classifier independent of the naturally occurring class distribution or error cost. A good classifier must achieve high TP rate and at the same time less FP rate (14).

2.2 Review of Related Works

In this section, some of the studies that have been conducted related to anaemia disease using traditional statistical techniques are discussed.

One of the studies conducted on the prevalence of anaemia was that of Samson and Fikre (15) entitled "Correlates of anaemia among women of reproductive age in Ethiopia: evidence from Ethiopian DHS 2005". The objective of the study was to assess the correlates of anaemia among women of reproductive age in Ethiopia. A quantitative cross-sectional study was carried out based on the secondary data of the Ethiopian Demographic Health Survey (EDHS) 2005. They downloaded the EDHS 2005 data from Measure DHS website in SPSS format. After data cleaning was done, data on a total of 5963 women of reproductive age were chosen for the study. In addition, information on a wide-range of potential independent variables such as socio-demographic, economic, dietary intake, nutritional status, micronutrient supplementation history, breastfeeding history, maternity services utilization, family planning use, fertility history, etc were extracted.

The data analysis was done using SPSS. Frequencies, percentage, mean and standard deviation were used for the descriptive analysis. Independent sample t-test and One-way Analysis of Variance (ANOVA) were applied to compare mean blood hemoglobin level across different categories of the independent variables. The findings of the study indicate that the prevalence of anaemia was 27.4%. And rural residence, poor educational and economic status, 30-39 years of age and high parity were key factors predisposing women to anaemia. From this result the researchers concluded that family planning, economic and educational empowerment of women should be instated in combating anaemia.

A cross-sectional community-based study with analytic component was conducted by Haider (16) in nine of the 11 regions of Ethiopia to assess the magnitude of anaemia, deficiencies of iron and folic acid and compare the factors responsible for anaemia among anaemic and non-anaemic women to identify the potential causes of anaemia in apparently healthy-looking Ethiopian women of childbearing age (15-49 years).

The study revealed that one in every three women had anaemia and deficiency of folic acid while one in every two had iron deficiency, suggesting that deficiencies of both folic acid and iron constitute the major micronutrient deficiencies in Ethiopian women.

Haider and Pobocik (17) also conducted the first large nutrition study of a representative sample of women aged 15 to 49 years in Ethiopia. The association of anaemia to demographic and health variables was tested by chi-square and a stepwise backward logistic regression model was applied to test the significant associations observed in chi square tests. The study indicated that intake of vegetables less than once a day and meat less than once a week was common and was associated with increased anaemia, and nutrition related and chronic illnesses are the most common causes of anaemia. The researchers concluded that moderate nutritional anaemia in the form of iron deficiency anaemia is a problem in Ethiopia and therefore, the need for improved supplementation to vulnerable groups is warranted to achieve the United Nation's Millennium Development Goals.

Sanku et al (35) conducted a study entitled “Prevalence of anaemia in women of reproductive age in Meghalaya: a logistic regression analysis”. The aim of the study was to determine the prevalence of anaemia among ever-married women of reproductive ages from the state Meghalaya, India, and to explore some factors commonly associated with anaemia. In the study, socioeconomic differentials are also presented to understand the prevalence of

anaemia. The study population consisted of 3934 ever-married women of reproductive ages (15-49 years) which were taken from the third Indian National Family Health Survey in 2005-2006. To explore the predictors responsible for the prevalence of anaemia by using different background characteristics such as age, place of residence, nutritional status, number of children ever born, pregnancy status, educational achievement, and economic status were used. As a response variable, anaemia levels were categorised as a dichotomous variable, and the predicted probabilities were worked out through a binary logistic regression model, to assess the contribution of the predictors on anaemia.

The findings of the study revealed that all the predictors, except total children ever born, were found to be statistically significant, and as a result of the mean haemoglobin concentration of 49.6% of the women were found to be anaemic. They also observed that women of the age group 20-24 years were at high risk of anaemia. Finally, the researchers conclude that pregnant, under nutritious, and poorest women are at high risk of anaemia,

Bentley and Griffiths (36) studied the burden of anaemia among women in India. They used The National Family Health Survey 1998/99 which provides nationally representative cross-sectional survey data on women's haemoglobin status, body weight, and diet, social, demographic and other household and individual level factors. The purpose of the study was to investigate the prevalence and determinants of anaemia among women in Andhra Pradesh. Their findings were prevalence of anaemia was high among all women, and poor urban women had the highest rates and odds of being anaemic.

2.3 Review of Research Papers

In this section, some of the researches that have been done by using data mining techniques are discussed.

One of the research conducted using data mining technique was that of Sanap et al. (37) entitled "Classification of Anaemia Using Data Mining Techniques". The study presented an analysis of the prediction and classification of anaemia in patients using data mining techniques. The aim of the study was to investigate C4.5 decision tree algorithm and SMO (Sequential Minimal Optimization) in WEKA, and to find which method gives most suitable technique for prediction and best possible classification of anaemia using CBC data and

generate decision tree. The dataset which contains 514 instances were constructed from complete blood count tests which are performed by collecting blood samples from 191 healthy individuals and 323 patients having disorders of anaemia.

After experimenting, they found that C4.5 decision tree algorithm classify anaemia perfectly with accuracy of 99.42%, and support vector machine with accuracy of 88.13%. From the findings they observed that C4.5 algorithm has best performance with highest accuracy and F-mesure.

Biset (38) for his master thesis used data mining techniques to predict low birth weight on Ethiopian Demographic and Health Survey data sets. The goal of the study was to predict low birth weight using EDHS 2005 (Ethiopia Demographic Health Survey) data set so as to build a model using data mining technique addressing the factors associated with low birth weight. He applied CRISP-DM methodology. J48 decision tree classifier and PART rule induction algorithms were selected for experiments.

The researcher compared the classification performance of the decision trees with tree pruning and without tree pruning, and found that tree pruning can significantly improve decision tree's classification performance. In general, the results from the study as the researcher described were encouraging which can be used as decision support aid for health practitioners. He further indicated that the extracted rules in both the algorithms are very effective for the prediction of low birth weight. Finally, from both algorithms he observed that attributes such as antenatal visits during pregnancy (antenatal care for pregnancy), mother's educational level, and marital status, iodine contents in salt, region, and age of mother, numbers of birth order and wealth index as well as place of residence are the most determinant factors to predict low birth weight.

Abel (39) also applied data mining techniques to design a predictive model for heart disease detection. The researcher's aim was to design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography. Knowledge Discovery in Database (KDD) methodology consisting of nine iterative and interactive steps was adopted.

He used a total of 7,339 records for experiments with implementation of J48 Decision Tree classifier, Naïve Bayes classifier and Neural Network. Experimental results revealed that all the models built using the above three techniques have high classification accuracy and are generally comparable in predicting heart disease cases. However, comparison that is based on True Positive Rate suggests that the J48 model performs slightly better in predicting heart disease with classification accuracy of 95.56%. Based on the findings, he concluded that data mining techniques can be used efficiently to model and predict heart disease cases, and the outcome of the study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease.

Muluneh (40) conducted a study by applying data mining in an attempt to investigate the potential applicability of data mining techniques in exploring the prevalence of diarrheal disease using the data collected from the diarrheal disease control and training centre of African sub Region II in Tikur Anbessa Hospital. He used the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology. Two machine learning algorithms from WEKA software, J48 Decision Trees(DT) and Naïve Bayes(NB) classifiers, were implemented to classify diarrheal disease records on the basis of the values of attributes 'Treatment' and 'Type of Diarrhea'.

Results of the experiments showed that J48 decision tree classifier has better classification and accuracy performance as compared to Naïve Bayes classifier. The researcher concluded that the study has proved that data mining techniques are valuable to support and scale up the efficacy of health care services provision process.

Elias (41) conducted a study on HIV status predictive modelling using data mining technology. The objective was to construct HIV status predictive model in support of scaling up of HIV testing in Addis Ababa. He used the CRISP-DM methodology for HIV status predictive modeling and discovering association rules between HIV status and selected attributes. In the study, J48 and ID3 algorithms were used to build and evaluate the models while apriori algorithm was used to discover association rules. WEKA 3.6 was used as the data mining tool. The findings of the researcher revealed that pruned J48 classifier that

predicts HIV status with 81.8% accuracy was developed, and association rule mining has also its potential in discovering the relationships of the selected attributes and HIV status.

Selam (42) used data mining technology to design a predictive model that can help predict the occurrence of measles outbreaks in Ethiopia. She used a hybrid six-step Cios KDP methodology, which consists of six basic steps such as: problem domain understanding, data understanding, data preparation, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge. Naïve Bayes and decision tree data mining techniques were employed to build and test the models using a dataset of 15631 records. Experimental results of the study revealed that among the implemented algorithms J48 has shown better prediction accuracy than Naïve Bayes classifier. From the results, she proved that applying data mining techniques on measles surveillance data to build a model that predicts the occurrence of measles outbreak in different Ethiopian Regions was possible.

Shegaw (43) applied data mining technology to predict child mortality patterns upon community-based epidemiological datasets collected by the Butajira Rural Health Project (BRHP) epidemiological study. The methodology which was used by the researcher had three basic steps. These were collecting of data, data preparation and model building and testing. BrainMaker and See5 software's were employed by the researcher so as to build models using neural network and decision tree techniques respectively.

He found that both the techniques yield comparable results for misclassification rates. However, unlike the neural network models, the results obtained from decision tree models provided simple and understandable rules that can be used by any health care professionals to identify cases for which the rule is applicable. In fact, the accuracy obtained from decision tree models also outperforms neural networks. He also found that best models to be using neural network technique by modifying default parameters of the program.

A study conducted by Teklu (44) has attempted to investigate the application of data mining techniques on Anti Retroviral Treatment (ART) service with the purpose of identifying the determinant factors affecting the termination/continuation of the service. This study applied classification and association rules using, J48 and apriori algorithms respectively, on 18740 ART patients' datasets. The methodology employed to perform the research work is CRISP-

DM. Finally the investigator proved the applicability of data mining on ART by identifying those factors causing the continuation or termination of the service.

Tesfahun (45) also attempted to construct adult mortality predictive model using data mining techniques so as to identify and improve adult health status using BRHP open cohort database. He used the hybrid model that was developed for academic research. Decision tree and Naïve Bayes algorithms were employed to build the predictive model by using a sample dataset of 62,869 records of both alive and dead adults through three experiments and six scenarios. The investigator found out that as compared to Bayes, the performance of J48 pruned decision tree reveals 97.2% of accurate results for developing classification rules that can be used for prediction. He further pointed out that if no education in family and the person is living in rural highland and lowland, the probability of experiencing adult death was 98.4% and 97.4% respectively with concomitant attributes in the rule generated. Finally, he suggested that education plays a considerable role as a root cause of adult death, followed by outmigration, and further comprehensive and extensive experimentation is needed to substantially describe the loss experiences of adult mortality in Ethiopia.

After reviewing the above literatures the researcher was motivated to work this study since no previous researches have been done by applying data mining techniques to predict the status of anaemia in Ethiopia using survey dataset. Hence this research has a great contribution to obtain useful information that can help health practitioners to develop a better strategy for preventing and treating anaemia using data mining technique.

CHAPTER THREE

BUSINESS UNDERSTANDING AND BACKGROUND INFORMATION

In this chapter, an attempt is made to present the concepts of the study problem, i.e. anaemia disease and background information on the survey dataset used for this study.

3.1 An Overview of Anaemia

3.1.1 Definition of Anaemia

Anaemia is a medical condition in which the red blood cell count or haemoglobin is less than normal. According to Sherman (3), anaemia is defined as a reduction in the number of circulating red blood cells, the haemoglobin concentration, or the volume of packed red cells (hematocrit) in the blood. Red Blood Cells (RBC) are the most numerous cells in the blood; approximately 20 billion of them circulate in the blood of an adult. They are required to transport oxygen to the tissues and organs of the body. RBC contains haemoglobin, an iron-containing protein that acts in the transportation of oxygen to the tissues and carbon dioxide from the tissues (5). Sherman (3) writes that since haemoglobin normally carries oxygen from the lungs to tissues, anaemia leads to the lack of oxygen in organs; and as all human cells depend on oxygen, varying degrees of anaemia can have a wide range of clinical consequences.

As given by WHO (1), anaemia status of the study population was graded according to cut-off points for diagnosis of anaemia. In other words, anaemia among different groups of the population was classified using the WHO haemoglobin reference definitions. Based on WHO haemoglobin threshold, therefore, a haemoglobin concentration of <120g/l and/or <12.0g/dl in women are the international cut-offs for any anaemia. On diagnosis of anaemia, according to WHO (1) recommended cut-offs, the following operational definitions were applied in women of reproductive age:

- Severe anaemia: Women with <70g/l and/or <7.0 g/dl of blood haemoglobin level.
- Moderate anaemia: Women with 70-99g/l and/or 7.0-9.9 g/dl of blood haemoglobin level.
- Mild anaemia: Women with 100 -120g/l and/or 10.0-12.0 g/dl blood haemoglobin level.

3.1.2 Main Causes of Anaemia

Commonly, anaemia is the final outcome of nutritional deficiency of iron, folate, vitamin B12, and some other nutrients (4). Many other causes of anaemia have also been identified. They include malaria, haemorrhage, infection, genetic disorders (hemoglobinopathies), parasite infestation (hookworm), chronic disease, and others.

Nutritional deficiency, due primarily to a lack of bio-available dietary iron, accounts for the majority of anaemia cases in the world (46). The contribution of other causes of anaemia depends on many factors, including level of economic development, climate, condition of health care, and the existence of anaemia control and prevention programs.

In semi-developed and developed countries, iron deficiency is the main cause of anaemia among women and children. In approximately 50 to 80 percent of anaemia cases, iron deficiency is considered to be the main etiologic factor (5). In many countries, however, a number of other factors besides iron deficiency contribute to the burden of anaemia. Of particular importance in developing countries are malaria and intestinal parasites, especially hookworm infestation. The level of contribution of these factors to the overall prevalence of anaemia depends on the magnitude of malaria epidemics, the existence of iron supplementation and fortification programs, and other conditions in each particular country.

Recently, the role of HIV epidemics as an important factor contributing to anaemia in countries of sub-Saharan Africa has been emphasized. It has been shown that HIV negatively affects the release of erythropoietin, which is a kidney hormone that stimulates production of red blood cells. Perhaps because of this mechanism, HIV-infected women, even without clinical symptoms of opportunistic infections, are more likely to become anaemic than HIV-free women (1).

3.1.3 Health consequences

Anaemia is an indicator of both poor nutrition and poor health. The consequences of anaemia for women include increased risk of low birth-weight or prematurity, prenatal and neonatal mortality, inadequate iron stores for the newborn, increased risk of maternal morbidity and

mortality, and lowered physical activity, mental concentration, and productivity (2). Women with even mild anaemia may experience fatigue and have reduced work capacity (7).

3.1.4 Classification of anaemia as a public health problem

The prevalence of anaemia as a public health problem is categorized as follows: <5%, no public health problem; 5–14.9%, mild public health problem; 15–39.9%, moderate public health problem; >=40%, severe public health problem (2). On a population level, anaemia prevalence can be distinguished as low, medium or high according to the level of prevalence suggested by WHO (2).

3.1.5 Anaemia and Maternal Health

Anaemia is known to have detrimental health implications, particularly for mothers and young children. Women with severe anaemia can experience difficulty meeting oxygen-transport requirements near and at delivery, especially if significant haemorrhaging occurs. This may be an underlying cause of maternal death and prenatal and prenatal infant loss (46).

In fact, unfavourable pregnancy outcomes have been reported to be more common in anaemic mothers than in non-anaemic mothers (1). The adverse effects of mild anaemia are less well documented than the effects of severe anaemia. However, in several studies, premature delivery, placental hypertrophy, and reduced excretion of estriol (maternal hormone) have been observed to be more common in mildly anaemic mothers than in non-anaemic mothers (4).

3.2 Ethiopia Demographic and Health Survey (EDHS) dataset

Ethiopia has made an effort to generate reliable demographic data by conducting a number of surveys. These include the 1981 Demographic Survey, the 1990 National Family and Fertility Survey (NFFS), the 1995 Fertility Survey of Urban Addis Ababa, and the 2000, 2005, and 2011 Ethiopia Demographic and Health Surveys (EDHS). The 1990 NFFS was the first nationally representative survey to yield substantial information on fertility, family planning, contraceptive use, and related topics. In addition to the topics covered by the NFFS, the 2000,

2005, and 2011 EDHS surveys collected information on maternal and child health, nutrition and breast-feeding practices, and HIV and other sexually transmitted diseases (8).

These three Demographic and Health Survey (DHS) in Ethiopia have been conducted at five years intervals by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health (9). The first ever Demographic and Health Survey (DHS) in Ethiopia was conducted in 2000. The second and the third EDHS was conducted in 2005 and 2011 respectively as part of the worldwide MEASURE DHS project (8,9).

As the source of data for this research is the 2011 EDHS, it is worth providing here brief background information on 2011 EDHS in relation with the prevalence of anaemia in women of reproductive age.

The 2011 Ethiopian Demographic and Health Survey (EDHS) were conducted under the aegis of the Ministry of Health and were implemented by the Central Statistical Agency from September 2010 through June 2011 with a nationally representative sample of nearly 18,500 households. It is the third comprehensive survey conducted in Ethiopia as part of the worldwide Demographic and Health Survey project, and is the second survey presenting results on HIV and anaemia prevalence. In general, the 2011 EDHS dataset has eight files. However, for the purpose of this study, the researcher used only those three files namely, Household, Individual, and Couples'.

The principal objective of the 2011 Ethiopia Demographic and Health Survey (EDHS) is to provide current and reliable data on fertility and family planning behaviour, child mortality, adult and maternal mortality, children's nutritional status, use of maternal and child health services, knowledge of HIV/AIDS, and prevalence of HIV/AIDS and anaemia. The specific objectives are these:

- Collect data at the national level that will allow the calculation of key demographic rates;
- Analyse the direct and indirect factors that determine fertility levels and trends;
- Measure the levels of contraceptive knowledge and practice of women and men by family planning method, urban-rural residence, and region of the country;

- Collect high-quality data on family health, including immunisation coverage among children, prevalence and treatment of diarrhoea and other diseases among children under age five, and maternity care indicators, including antenatal visits and assistance at delivery;
- Collect data on infant and child mortality and maternal mortality;
- Obtain data on child feeding practices, including breastfeeding, and collect anthropometric measures to assess the nutritional status of women and children;
- Collect data on knowledge and attitudes of women and men about sexually transmitted diseases and HIV/AIDS and evaluate patterns of recent behaviour regarding condom use;
- Conduct haemoglobin testing on women age 15-49 and children 6-59 months to provide information on the prevalence of anaemia among these groups;
- Carry out anonymous HIV testing on women and men of reproductive age to provide information on the prevalence of HIV.

This information is essential for informed policy decisions, planning, monitoring, and evaluation of programmes on health in general and reproductive health in particular at both the national and regional levels. A long-term objective of the survey is to strengthen the technical capacity of the Central Statistical Agency to plan, conduct, process, and analyse data from complex national population and health surveys.

Moreover, the 2011 EDHS provides national and regional estimates on population and health that are comparable to data collected in similar surveys in other developing countries and to Ethiopia's two previous DHS surveys, conducted in 2000 and 2005. Data collected in the 2011 EDHS add to the large and growing international database of demographic and health indicators.

The sample for the 2011 EDHS was designed to provide population and health indicators at the national (urban and rural) and regional levels. The sample design allowed for specific indicators, such as contraceptive use, to be calculated for each of Ethiopia's 11 geographic/administrative regions (the nine regional states and two city administrations). The 2007 Population and Housing Census, conducted by the CSA, provided the sampling frame from which the 2011 EDHS sample was drawn.

Administratively, regions in Ethiopia are divided into zones, and zones, into administrative units called weredas. Each wereda is further subdivided into the lowest administrative unit, called kebele. During the 2007 census each kebele was subdivided into census enumeration areas (EAs), which were convenient for the implementation of the census. The 2011 EDHS sample was selected using a stratified, two-stage cluster design, and EAs were the sampling units for the first stage. The sample included 624 EAs, 187 in urban areas and 437 in rural areas.

Households comprised the second stage of sampling. A complete listing of households was carried out in each of the 624 selected EAs from September 2010 through January 2011. Sketch maps were drawn for each of the clusters, and all conventional households were listed. The listing excluded institutional living arrangements and collective quarters (e.g., army barracks, hospitals, police camps, and boarding schools). A total of 17,817 households were selected for the sample, of which 17,018 were found to be occupied during data collection. Of these, 16,702 were successfully interviewed, yielding a household response rate of 98 percent.

In the interviewed households 17,385 eligible women were identified for individual interview; complete interviews were conducted for 16,515, yielding a response rate of 95 percent. Similarly, a total of 15,908 eligible men were identified for interview; completed interviews were conducted for 14,110, yielding a response rate of 89 percent. In general, response rates were higher in rural areas than urban areas, for both women and men.

The 2011 EDHS used three questionnaires: the Household Questionnaire, the Woman's Questionnaire, and the Man's Questionnaire. These questionnaires were adapted from model survey instruments developed for the MEASURE DHS project to reflect the population and health issues relevant to Ethiopia. Issues were identified at a series of meetings with the various stakeholders. In addition to English, the questionnaires were translated into three major languages—Amharigna, Oromiffa, and Tigrigna.

The Household Questionnaire was used to list all the usual members and visitors of selected households. Basic information was collected on the characteristics of each person listed,

including age, sex, education, and relationship to the head of the household. For children under age 18, survival status of the parents was determined. The data on the age and sex of household members obtained in the Household Questionnaire were used to identify women and men who were eligible for the individual interview.

The Woman's Questionnaire was used to collect information from all women age 15-49. These women were asked questions on the following topics:

- Background characteristics such as age, education and media exposure
- Birth history and childhood mortality
- Knowledge and use of family planning methods
- Fertility preferences
- Antenatal, delivery and postnatal care
- Breastfeeding and infant feeding practices
- Vaccinations and childhood illnesses
- Marriage and sexual activity
- Women's work
- Husband's background characteristics
- Awareness and behaviour regarding AIDS and other sexually transmitted infections (STIs)
- Adult mortality, including maternal mortality

The 2011 EDHS included height and weight measurement, anaemia testing, and blood sample collection for HIV testing in the laboratory. Blood specimens were collected for anaemia testing from all children age 6-59 months, women age 15-49, and men age 15-59 who voluntarily consented to the testing. Blood samples were drawn from a drop of blood taken from a finger prick (or a heel prick in the case of young children with small fingers) and collected in a microcuvette.

Haemoglobin analysis was carried out onsite using a battery-operated portable HemoCue analyser. Results were given verbally and in writing. Parents of children with a haemoglobin level under 7 g/dl which is severe anaemia according to WHO cut-off were instructed to take the child to a health facility for follow-up care. Likewise, non-pregnant women were referred for follow-up care if their haemoglobin level was below 7 g/dl, and pregnant women and men

were referred if their haemoglobin level was below 9 g/dl which is moderate anaemia according to WHO cut-off. All households in which anaemia testing was conducted received a brochure explaining the causes and prevention of anaemia.

According to the 2011 EDHS (8), seventeen percent of Ethiopian women ages 15-49 years in all selected households are anaemic. A higher proportion of pregnant were anaemic than women who are breastfeeding and women who are neither pregnant nor breastfeeding. Anaemia prevalence also varies by urban and rural residence, a higher proportion of women in rural areas are anaemic than those in urban areas.

Also, women in Somali, Affar, and Dire Dawa regions have relatively high prevalence of anaemia. Women in Addis Ababa and the SNNP and Tigray regions are at the other end of the range, with relatively low prevalence of anaemia. Women with no education are twice as likely to be anaemic as women with more than secondary education. Similarly, anaemia prevalence decreases as wealth status increase (8).

CHAPTER FOUR

RESEARCH DESIGN AND DATA PREPARATION

In this study, as it has been mentioned in section 1.4, the KDD methodology was followed to explore the application of data mining techniques on the obtained dataset. In this chapter, therefore, the five consecutive steps of KDD as shown in Figure 4.1 have been reviewed in detail. The researcher starts by introducing the raw data and its related processing, and then explains attribute selection, attribute reduction, and other pre-processing steps taken to reach to the pre-processed data. Finally, the data mining and evaluation steps were discussed.

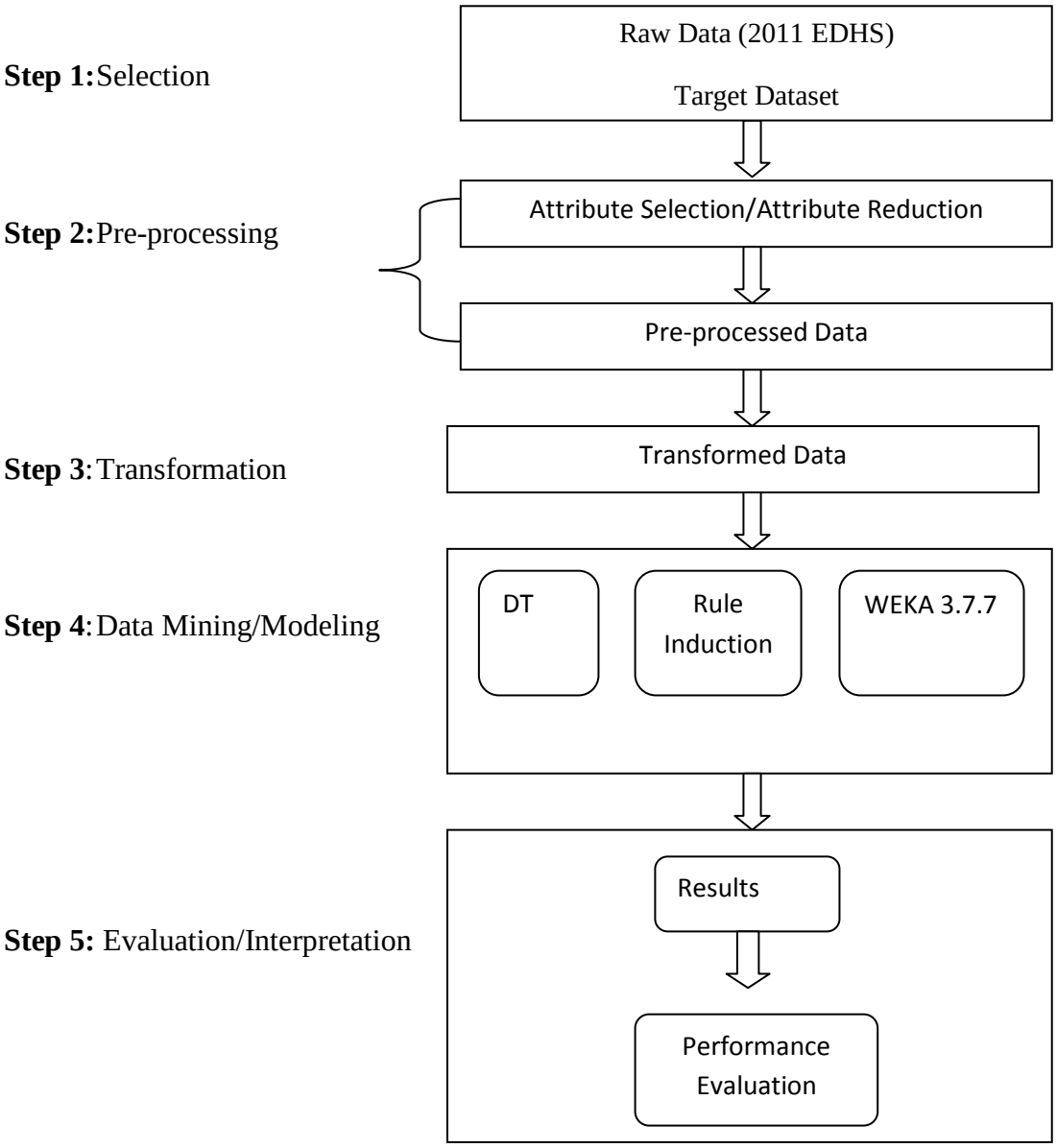


Figure 4.1: Diagrammatic Overview of the Research Design

The following is the description of each of the steps.

4.1 Data Selection

Data selection method is the first step which includes raw data extraction and selecting the target dataset that are used for the study problem. In a case where we have millions of data points, we have to select a subset of the database to perform the required knowledge discovery steps. Therefore, selection is the process of choosing on the right data from the database on which the tools in data mining can be used to extract useful information, knowledge and pattern from the provided raw data (48).

As it has been stated under section 1.3 of the first chapter, the aim of this study is to predict the status of anaemia in women aged 15-49 by applying data mining techniques using the 2011 Ethiopia Demographic and Health Survey (EDHS) survey datasets. And the source data employed for this research was those three files of the 2011 EDHS datasets indicated in section 1.4.2. The 2011 EDHS which is a nationally representative survey dataset that includes women aged 15-49, children aged 6-59 months, and men aged 15-59 was originally stored in SPSS format. This dataset was organized in rows and columns where each column represents an attribute (variable) and each row stands for a single record of an individual. In the source dataset a total of 920 attributes (columns) and 45,539 records (rows) were identified.

The researcher took a zipped copy of this original data set and all the necessary operations of data selection were carried out on this data set. After collecting the SPSS format of the original survey dataset, Microsoft Excel is used for selecting the target data set required for this research work. In this initial step, from the whole dataset, data only on women aged 15-49 years was extracted to achieve the objectives of this study. Thus, using the selection facility provided by Microsoft Excel 2007 software, 6697 records on women aged 15-49 have been selected.

4.2 Data Pre-processing

Once the target data that was used for the knowledge discovery is determined, the data is then pre-processed to ensure the reliability of the data. In this step, in addition to data pre-

processing/data preparation, tasks that include attribute selection, attribute reduction, and data cleaning particularly handling missing values are undertaken.

According to Witten and Frank (24), raw data cannot be used directly for processing with the machine learning algorithms. They first need to be pre-processed into machine understandable format. Raw data can be stored in several formats, including text, Excel or other database types of file. Sometimes the raw data is not in any format. Having data already in a format understandable by algorithms can result in better time efficiency with respect to processing of the data (24). Therefore, each algorithm requires data to be submitted in a separate format.

In this study, the raw data obtained from the 2011 EDHS dataset were initially in SPSS format which is not understandable by the WEKA tool. To make such data format understandable by WEKA, the researcher saved the exported SPSS file format into Microsoft Excel as a CSV (Comma Separate Values) file format to make ready for an input to WEKA 3.7.7 data mining tool. The last task done to make the data format suitable for WEKA tool was converting the CSV file into ARFF (Attribute-Relation File Format) file format.

4.2.1 Attribute Selection

As it is mentioned in section 4.1, the original dataset from which the target dataset for this research work has been selected consisted of a total of 920 attributes. However, all these attributes found in the original dataset were not necessary for this study. Thus, only relevant attributes were considered for the specific learning task to achieve the objectives of this research work.

According to Kim and Menczer (49), the rationale behind attribute or feature selection which is considered as a pre-processing stage is to choose a subset of input variables by removing attributes with little or no predictive information with the objectives of increasing learning accuracy.

It is intuitively reasonable from these objectives to think that large number of attributes or features is not informative because they are either irrelevant or redundant to the prediction (49). Although there are different methods that can be applied to feature/attribute selection,

the researcher used the manual way of selecting attributes due to its emergence as an alternative approach to attribute selection.

In first iteration, based on insights taken from literature review, domain experts knowledge from public health field particularly nutritionist whose list is shown in ANNEX 1, and an extensive review of all attributes detailed in 2011 EDHS documentation files, 25 potential independent attributes are selected based on their background characteristics which could determine the status of anaemia in women of reproductive age.

These attributes are categorized under four modules: Socio-demographic (10 attributes), nutritional status (2 attributes), reproductive health related (9 attributes), and health behaviour (4 attributes). The researcher further analyzed each of these module attributes and introduces the selected set of attributes in each category that are finally used in the modelling phase.

4.2.1.1 Socio-demographic attributes

Table 4.1 summarizes the titles and descriptions of 10 socio-demographic attributes as shown in the 2011 EDHS dataset.

Table 4.1: Socio-Demographic attributes

No.	Attribute Name	Descriptions
1	AGE	Women's age in years
2	REGION	Region where the survey conducted
3	RESIDENCE	Type of place of residence
4	EDUCLEV	Highest educational level
5	EDUCATT	Educational attainment
6	RELIGION	Respondent's religion
7	SEX	Sex of household
8	WEALTH	Respondent's wealth index
9	MARRY	Current marital status
10	EMPLOY	Respondent's current working or employment status

All these attributes are categorical (either flag, ordinal or nominal), except AGE attribute which is continuous (scale) variable.

AGE represents the age data of women tested for anaemia. In the 2011 EDHS, it has been top-coded to 49 in original data and ranges from 15 to 49. REGION indicates the census region of each living unit including the survey participant. It accepts values of 1 to 11 for Tigray, Afar, Amhara, Oromyia, Somali, Benshangul-Gumuz, SNNP, Gambella, Harari, Addis Ababa, and Dire Dawa respectively. RESIDENCE represents the type of place of respondents' residence, either urban or rural, with a value equal to 1 or 2 respectively.

EDUCLEV is the highest educational level of respondents that has been top-coded to 3, where 0 shows no education, 1 shows primary level, 2 shows secondary, and 3 stand for higher level. EDUCATT indicates the educational attainment of households. It accepts values of 1 to 5 for no education, incomplete primary, complete primary, incomplete secondary, complete secondary, and higher respectively.

RELIGION indicates the religion of the person being interviewed and includes orthodox, catholic, protestant, muslim, traditional or other designated by the values of 1 to 6 respectively. SEX indicates the gender of each person, either male or female, with a value equal to 1 or 2. WEALTH is the categorical variable that shows the wealth status of the respondents. It accepts values of 1 to 5 for poorest, poorer, middle, richer, or richest respectively. MARRY shows the marital status of the person in the month of the survey, and top-coded to 5, where 0 for never in union, 1 for married, 2 for living with partner, 3 for widowed, 4 for divorced, and 5 stands for no longer together/separated respectively. EMPLOY represents the current working or employment status of the respondents. It is a flag attribute accepting 0 (No) or 1 (Yes) values.

Bearing in mind the manual way of selecting relevant attribute from the original dataset, and on the basis of discussion with domain experts, some of the attributes removed from the socio-demographic set are EDUCATT, SEX, and RELIGION.

EDUCATT attribute which indicates the educational attainment of the respondents is removed because of having redundant information with EDUCLEV attribute which contains more relevant information for this particular study. SEX attribute is removed due to its irrelevance to this specific study since the problem domain of this study is women of

reproductive age. RELIGION is removed since it is a sensitive attribute to describe the study variable. Therefore, the final attributes in the socio-demographic set are AGE, REGION, RESIDENCE, EDUCLEV, WEALTH, MARRY, AND EMPLOY.

4.2.1.2 Nutritional status attributes

Table 4.2 summarizes the titles and descriptions of 2 nutritional status attributes as shown in the 2011 EDHS dataset.

Table 4.2: Nutritional status attributes

No.	Attribute Name	Descriptions
1	BMI	Body mass index of a person
2	IRON	Iron supplementation for women of the reproductive age

Among these two attributes IRON attribute is categorical, which is flag attributes accepting 0 (No) or 1 (Yes) values whereas BMI is a continuous (scale) variable. BMI is a measure of body mass index and person's height and weight (8).

According to WHO (6) definitions, BMI in adults ranges from low body mass index (<18.5), normal weight (18.5 – 24.9) inclusive, and overweight (>=25.0). Based on insights taken from literature review in which the nutritional status is considered as an important determinants for the prevalence of anaemia, the above two attributes, namely, BMI and IRON are includes in the final list for the nutritional status attributes.

4.2.1.3 Reproductive health related attributes

Table 4.3 summarizes the titles and descriptions of 9 reproductive health related attributes as shown in the 2011 EDHS dataset.

Table 4.3: Reproductive health related attributes

No.	Attribute Name	Descriptions
1	BIRTHS	Give births in the months of interview
2	PREGNANCY	Current pregnancy status

3	TERMPREGNANCY	Terminated pregnancy status
4	CONTRACEPTIVE	Current contraceptive use
5	HEALTHF	Visited health facility in last 12 months
6	BREASTF	Current breastfeeding status
7	DELIVERY	Place of delivery
8	ANC	Antenatal care received by women
9	POSTNATAL	Check-up after delivery

All these attributes are categorical with a flag attribute accepting 0 (No) or 1 (Yes) values, except DELIVERY with nominal variable accepting 1 to 2 for home or health institutions respectively. Due to their relevance to this particular study, all these attributes are included in the final attribute list for reproductive health related attributes.

4.2.1.4 Health behaviour attributes

Table 4.4 summarizes the titles and descriptions of 4 health behaviour attributes as shown in the 2011 EDHS dataset.

Table 4.4: Health behaviour attributes

No.	Attribute Name	Descriptions
1	SMOKECIG	Smoke cigarettes
2	SMOKEOTH	Smokes other such as suret, shisha, and gaya
3	CHAT	Ever chewed chat
4	ALCHOL	Ever took an alcoholic drink

All health behaviour attributes are of flag type accepting 0 (No) or 1 (Yes) values. On the basis of their relevance, all these attributes are also included in the final attribute list for health behaviour attributes.

4.2.2 The missing values

Missing attributes values in the data is most likely associated with unavailability of interesting information, lack of knowhow on the importance of data at the time of entry, misunderstanding of the data, the respondents him/her self may refuse to answer certain questions or they may not know the answer exactly or may answer in an unexpected manner (11).

For many real-world applications of data mining, a number of problems are faced while bringing the data into proper format. Missing data is the most common problem that comes up during the data analysis process. Missing values minimizing the accuracy of classification and rules generated by the selected data mining algorithm. Missing values lead to the difficulty of extracting useful information from that data set.

As it is suggested by Kantardzic (11), it is possible to ignore the attributes with too many missing values or more than 50% missing values. Avoiding the missing data is not time consuming and at the same time it is very easy to follow. But there are many drawbacks associated with this method. Deleting records may result in losing some information (11). If the sample data size is large avoiding some records or attributes may not affect the results, but still we need to keep in mind we are losing something important if we ignore attributes with too many missing values without considering all aspects of the problem under study.

In this study, as it is indicated in section 1.4.2, three files of the 2011 EDHS dataset such as Household, Individual, and Couples' which contained a total of 45, 359 records and 920 attributes were employed as a source data. However, while looking each file, some attributes values which are presented in one file missed in another files. For instance, in the Household file, attributes such as ANC, POSTNATAL, DELIVERY, and IRON had more than 50% missing values, while they have less than 50% missing values in other files mentioned above. Thus, to manage such problems of missing values, the researcher followed the data imputation method to substitute the missing values rather than simply avoiding those attributes which have more than 50% missing values.

After using such way of substituting missing values particularly for those attributes which initially had more than 50% missing values, therefore, in the final dataset of this research

work, 12 out of 22 attributes have missing values. Table 4.5 shows the value labels for missing values in each pertinent field.

Table 4.5: Missing values in the final dataset

No.	Attribute Name	Missing Value	Missing %
1	TERMPREGNANCY	9	0.1
2	HEALTHF	11	0.16
3	BMI	152	2.7
4	SMOKECIG	2	0.02
5	SMOKEOTH	11	0.16
6	EMPLOY	1	0.01
7	DELIVERY	989	14.7
8	POSTNATAL	1032	15.4
9	ANC	1410	21.0
10	IRON	628	10.3
11	CHAT	2	0.02
12	ALCHOL	3	0.03

In order to handle the problem of attributes with missing values in the dataset created for a given data mining task, since all attributes shown in Table 4.5 have below 50% missing values, the researcher employed the use of modal value approach for substituting the missing after identifying their attribute type.

Based on the method, in the final dataset of this study, since all attributes with missing values are nominal categorical variables, for any missing value in such field, the modal (most frequent) value was substituted. The modal values for the above nominal attributes are shown in Table 4.6.

Table 4.6: Modal values representing missing values

No.	Attribute Name	Modal Value
1	TERMPREGNANCY	No
2	HEALTHF	No
3	BMI	Normal
4	SMOKECIG	No
5	SMOKEOTH	No
6	EMPLOY	No
7	CHAT	No
8	ALCHOL	No
9	DELIVERY	Home
10	POSTNATAL	No
11	ANC	Yes
12	IRON	No

4.2.3 Selected attributes in the final dataset

After attribute reduction efforts and handling missing values on primary set of 25 attributes, the final dataset has 22 attributes which are categorized in four modules:

- Socio-demographic (7 attributes): AGE, REGION, RESIDENCE, EDUCLEV, WEALTH, MARRY, AND EMPLOY.
- Nutritional status (2 attributes): BMI and IRON
- Reproductive health related (9 attributes): BIRTHS, PREGNANCY, TERMPREGNANCY, CONTRACEPTIVE, HEALTHF, ANC, POSTNATAL, DELIVERY and BREASTF.
- Health behavior (4 attributes): SMOKECIG, SMOKEOTH, CHAT, and ALCHOL

4.3 Attribute transformation

By completing preliminary processing of attributes, the final dataset has 22 variables for 6697 records. Before feeding such dataset into the software, attribute transformation including dimension reduction and attribute construction are conducted in some fields to make models work better.

In this research, WEALTH attribute is re-binned into a new nominal attribute with values of 1 to 3 for poor, middle, and rich respectively. It reduces the number of categories for WEALTH attribute from five to three and makes interpretation much easier. MARRY variable also re-binned into four new attribute which accepts two nominal values. The new nominal attributes of MARRY, accepts values of 1, 2, 3, and 4 for never married, married/living together, widowed, and divorced/separated, respectively.

With respect to nutritional status attributes, the only attribute which is the BMI attribute has been changed from continuous to ordinal values by using the same definitions used in the 2011 EDHS documentation that are comparable with the WHO definitions for body mass index in adults. The new BMI attribute accepts values from 1 to 3 in an ordinal basis for low, normal, and overweight respectively. Table 4.7 shows all 19 attributes along with each attribute descriptions, data type, and values they accept in the final dataset.

Table 4.7: The Final dataset attribute list – EDHS 2011 Dataset

No.	Attribute Name	Descriptions	Values	Data Type
1	AGE	Women's age in years	1,2,3,4,5,6,7	Nominal
2	REGION	Region where the survey conducted	1,2,3,4,5,6,7 ,8,9,10,11	Nominal
3	RESIDENCE	Type of place of residence	1,2	Nominal
4	EDUCLEV	Highest educational level	0,1,2,3	Nominal
5	WEALTH	Respondent's wealth index	1,2,3	Nominal
6	MARRY	Current marital status	1,2,3,4	Nominal
7	EMPLOY	Respondents current working or employment status	0,1	Flag
8	BMI	Body mass index of person	1,2,3	Ordinal

9	IRON	Iron supplementation for women of the reproductive age	0,1	Flag
10	BIRTHS	Give births in the month of interview	0,1	Flag
11	PREGNANCY	Current pregnancy status	0,1	Flag
12	TERMPREGNANCY	Terminated pregnancy status	0,1	Flag
13	CONTRACEPTIVE	Current contraceptive use	0,1	Flag
14	HEALTHF	Visited health facility in last 12 months	0,1	Flag
15	BREASTF	Current breastfeeding status	0,1	Flag
16	DELIVERY	Place of delivery	1,2	Nominal
17	ANC	Antenatal care received by women	0,1	Flag
18	POSTNATAL	Check-up after delivery	0,1	Flag
19	SMOKECIG	Smoke cigarettes	0,1	Flag
20	SMOKEOTH	Smokes other such as suret, shisha, and gaya	0,1	Flag
21	CHAT	Ever chewed chat	0,1	Flag
22	ALCHOL	Ever took an alcoholic drink	0,1	Flag

4.3.1 Target/Class attributes

With respect to class attribute, as the purpose of this research is to predict the status of anaemia among women of reproductive age, the target attribute selected in this study is anaemia level. This attribute initially has four different values such as mild, moderate, severe, and not anaemic. However, to make the interpretation of the result much easier, like other independent variables, it is re-binned into a new nominal attribute with values of 1 to 2 for any anaemic and not anaemic respectively. These new nominal attributes have been changed by using the same definitions of the threshold or cut-off point of the haemoglobin level adjusted for altitude used in the 2011 EDHS documentation that are comparable with the WHO definitions for the prevalence of anaemia in the world.

Generally, the target/class attribute selected in this study is anaemia level that has two nominal values, namely, any anaemic and not anaemic. This class attribute is considered as dependent variable while the rest of the variables indicated in Table 4.7 are the independent attributes for this particular study.

4.4 Data mining

Today, data mining is more popular in different settings including health care areas by mining knowledge from the massive repositories. The data mining step is the major step in KDD. This is when the cleaned and pre-processed data is sent into the intelligent algorithms for classification, clustering, similarity search within the data, and so on. Here we chose the algorithms that are suitable for discovering patterns in the data. Some of the algorithms provide better accuracy in terms of knowledge discovery than others. Thus selecting the right techniques and tools can be crucial at this point (24).

Usually the selection of modelling technique is based on the objective formulated. Since the purpose of this research is to predict the status of anaemia in women of reproductive age, the data mining techniques that has been selected is classification techniques. As a data mining tools, the researcher used WEKA software version 3.7.7 for reasons of accessibility and familiarity, and consists of different techniques and algorithms that have been used for building models.

There are a number of classification techniques that are applicable in data mining study. According to Han and Kamber (14), commonly used techniques for data classification and prediction in data mining are Decision Tree based Methods, Rule-based Methods; Memory based reasoning, Neural Networks, Naïve Bayes and Bayesian Belief Networks, and Support Vector Machines. However, for the purpose of this research only two classification algorithms, mainly, Decision Tree and Rule induction are selected.

These algorithms were selected due to their popularity and usefulness in solving data mining classification problems. Moreover, their advantages such as easy to interpret/ understand and visualize the result, fast at classifying unknown records/new instances, and easily handle missing values and numeric attributes compared to other classification techniques are taken as other reason for selecting only these two classification techniques.

4.4.1 J48 Decision Tree

The decision tree algorithm used in this research is J48 algorithm, which is one of the most common decision tree construction algorithms (24). It is WEKA's implementation of this decision tree learner. J4.8 actually implements a latter and slightly improved version called C4.5 revision 8, which was the last public version of this family of algorithms before the commercial implementation C5.0 was released (24).

J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, following a path from the root to the leaves of the tree, a sequence of such tests is performed resulting in a decision about the appropriate class of the dataset. J48 is a greedy algorithm i.e. it constructs trees in a top-down recursive or divide-and-conquer manner and orders the class rule sets so as to minimize the number of false-positive error (11).

4.4.2 PART Rule induction

The rule induction classifier used in this study is PART, which is one of the most commonly rule based classifications method. Rules are a good way of representing information or bits of knowledge (14). PART is usually called a separate-and-conquer technique because it identifies a rule that covers instances in the class (and excludes ones not in the class), separates them out, and continues on those that are left (24). In other words, it combines the divide-and-conquer strategy with separate-and-conquer strategy of rule learning. Such algorithms have been used as the basis of many systems that generate rules. The algorithm generates sets of rules called 'decision lists' which are ordered set of rules. PART obtains rules from partial decision trees using J48, and builds a partial C4.5 decision tree and converts the "best" leaf into a rule (24).

4.4.3 WEKA Tool

WEKA open software is one of the data mining tools that are used for data mining research purposes. In this research, WEKA software version 3.7.7 is employed. A number of data mining methods are implemented in the WEKA software. Some are based on decision trees like the J48 decision tree, and some are rule-based like PART and decision tables, and others

are based on probability and regression, like the Naïve Bayes algorithm (24). For the purpose of this research, the researcher used J48 decision tree and PART rule induction.

4.5 Performance Evaluation

Evaluation is the most important step in any data mining task and the key to making real progress in data mining, because it gives us a clear view regarding the strength of each model individually, and enables us to compare different models based on common performance measures. This step determines whether or not the models are working.

Performance of models is analyzed based on a series of measures. The standard performance measures that were used for evaluation of classifiers in this study include: confusion matrix such as correctness accuracy, sensitivity (TP Rate), specificity (TN Rate), precision, recall, F-measure, and ROC (Receiver Operating Characteristic). Each of these was used where appropriate in the analysis of the performances. Apart from the major performance criteria mentioned, the speed was also used as performance measures.

Speed refers to the execution time it takes a classifier to be trained. This translates to how demanding the algorithm is, in terms of the resources it will require to run. To measure this, the CPU time was used to assess how long it takes during building and training. This was recorded in seconds and computed. All the experiments were performed on Personal Computer with 4.00GB and 2.53GHz processor for comparability purposes.

Finally, after all of the KDD method processes are undertaken successfully, a report will be prepared by summarizing the research result and make available for dissemination.

CHAPTER FIVE

MODELING AND ANALYSIS OF THE RESULTS

In this chapter, the results of the modeling and performance evaluation phases are presented in four parts. First, we present the different experimentation conducted on J48 decision tree and PART rule induction to build the models with all 22 attributes and selected 13 attributes as indicated in chapter four. The second part presents the performance comparisons of the models. The comparison and evaluation were executed based on the results of the experiment performed using standard performance evaluation metrics. Here, the analysis and discussion of the output in the context of the problem were also presented. In the third part, we introduce rules extracted from the models. Finally, in the fourth part, we present the potential set of attributes which can reasonably predict any anaemic instances.

In this study, during all experiments, a 10-fold cross validation were executed to minimize the bias associated with the random sampling of the training and testing data samples through repeating the experiments ten times. In each experiment, all the default parameter settings already offered by the WEKA tool were used, except for the parameter "Unpruned" which had a default value "False" and was changed to "True" in order to observe the performance of the models. The algorithm was trained with the default settings of WEKA of these parameters as shown in ANNEX 2. The intention of executing experimentation here was to investigate the best experiment that produces the best model for predicting any anaemic instances.

5.1 Modeling

In this section, the results of the models built using the algorithms used in this study are presented. Basically, four experiments were conducted and as a result a total of eight models were developed. During the experiments, two categories considering four scenarios in each were carried out with the assumption of assessing the performance of models from different levels and perspectives. The first category of the experiments was conducted using all 22 attributes, whereas the second category was executed using selected 13 attributes.

To select such 13 attributes, the researcher used the information gain measure which was implemented by WEKA attribute ranking filter. Ranking the attributes basically indicates the

relative importance of each input attribute in making a prediction (24). In this regard, the 13 selected attributes using the WEKA ranker filter were shown in Table 5.1

Table 5.1: The Selected 13 attributes using WEKA ranker filter

No.	Attribute Name	Descriptions
1	AGE	Women's age in years
2	REGION	Region where the survey conducted
3	EDUCLEV	Highest educational level
4	MARRY	Current marital status
5	BMI	Body mass index of person
6	IRON	Iron supplementation for women of the reproductive age
7	BIRTHS	Give births in the month of interview
8	PREGNANCY	Current pregnancy status
9	TERMPREGNANCY	Terminated pregnancy status
10	CONTRACEPTIVE	Current contraceptive use
11	POSTNATAL	Check-up after delivery
12	CHAT	Ever chewed chat
13	ALCHOL	Ever took an alcoholic drink

5.1.1 Modeling using J48 decision tree

The modeling efforts start with building classification model using J48 decision trees. To build the models, two experiments considering four scenarios were conducted. The two scenarios contained all 22 input attributes and the other two contained the selected 13 attributes. The results obtained from these experiments were summarized in Table 5.2 with their respective performance measures.

Table 5.2: J48 decision tree experiment results – Final EDHS 2011 Dataset

Experiment	Model	Detailed Performance Measures							
		Accuracy	TP Rate	TN Rate	Precision	Recall	F-Measures	ROC area	Speed
1	J48 Pruned with all attributes	92.9%	88.3%	95%	0.891	0.883	0.887	0.968	0.23
	J48 Pruned with selected attributes	84.3%	71.5%	90.3%	0.772	0.715	0.742	0.907	0.11
2	J48 Unpruned with all attributes	95.6%	91.9%	95%	0.91	0.919	0.914	0.977	0.11
	J48 Unpruned with selected attributes	85.3%	76.8%	89.2%	0.767	0.768	0.767	0.924	0.05

As can be seen from Table 5.2, two experiments were executed. The first experiment was designed to evaluate the performance of J48 pruned classifier in predicting any anaemic women, whereas the second experiment was designed to evaluate the performance of J48 Unpruned model in predicting any anaemic women of reproductive age. In each experiment, two scenarios were considered and the first scenario of each experiment contained all 22 attributes and the second scenario contained selected 13 attributes.

In the first scenario of each experiment the model was run on a full training set containing 6,697 instances with all 22 attributes. In the first experiment, scenario one and two took 0.23 and 0.11 second, respectively, to build the model, while in the second experiment, scenario one and two took 0.11 and 0.05 second.

In modeling using J48 decision tree classifier, various results were obtained during the two experiments. The first experiment shows that the model built using J48 pruned with all attributes correctly classified 6222 (92.9%) instances, while 475 (7.1%) of the instances were classified incorrectly. The result of J48 pruned with selected attributes shows that the model

correctly classified 5648 (84.3%) instances, while 1049 (15.7%) of the instances were classified incorrectly.

The second experiment indicates that the model built using J48 Unpruned decision tree with all attributes correctly classified 6333 (94.6%) instances, while 364 (5.4%) of the instances were classified incorrectly. The result of J48 Unpruned model with selected attributes indicates that the model correctly classified 5713 (85.3%) instances, while 984 (14.7%) of the instances were classified incorrectly.

While considering the models built in each experiment, model's performance in both the pruned and unpruned scenarios when using all input attributes revealed a better performance than when using selected attributes with their respective detailed performance measures except the speed. However, the unpruned models performed better than the pruned ones. The experimental outputs of the J48 unpruned model with all attributes is presented in ANNEX 3.

With regards to the confusion matrix, which is a useful tool for analyzing how well the classifier can recognize tuples of different classes, the output summary of J48 decision tree are shown in Tables 5.3 and 5.4.

Table 5.3: Confusion Matrix of J48 decision tree for Experiment 1

Model 1	**Confusion Matrix**	
J48 pruned with all attributes	Any anaemic	Not anaemic
	1864 (88.2%)	248 (11.8%)
	227 (5%)	4358 (95%)
Model 2	**Confusion Matrix**	
J48 pruned with selected attributes	Any anaemic	Not anaemic
	1510 (71.5%)	602 (28.5%)
	447 (10%)	4138 (90%)

As it is shown in Table 5.3, the first model correctly classified 1864 (88.2%) women out of 2112 women as "Any anaemic" and the remaining 248 (11.8%) women were incorrectly

classified as “Not anaemic” while in fact they belong to “Any anaemic”. This model also correctly classified 4358 (95%) women as “Not anaemic” and the remaining 227 (5%) women were incorrectly classified as “Any anaemic” while they actually belong to “Not anaemic”. The second model also correctly classified 1510 (71.5%) women as “Any anaemic” and 4138 (90%) women as “Not anaemic”. The remaining 602 (28.5%) and 447 (10%) women were classified incorrectly as “Not anaemic” and “Any anaemic” while in fact they belong to “Any anaemic” and “Not anaemic”, respectively.

Table 5.4: Confusion Matrix of J48 decision tree for Experiment 2

Model 3	**Confusion Matrix**	
J48 Unpruned with all attributes	Any anaemic	Not anaemic
	1941(91.9%)	171(8.1%)
	193 (4.4%)	4392 (95.6%)
Model 4	**Confusion Matrix**	
J48 Unpruned with selected attributes	Any anaemic	Not anaemic
	1622 (76.8%)	490 (23.2%)
	494 (10.8%)	4091 (89.2%)

The third model, as depicted in Table 5.4, correctly classified 1941 (91.9%) women out of 2112 women as “Any anaemic” and the remaining 171 (8.1%) women were classified incorrectly as “Not anaemic” while in fact they belong to “Any anaemic”. The model also correctly classified 4392 (95.6%) women as “Not anaemic” and the remaining 193 (4.4%) women were classified incorrectly as “Any anaemic” while they actually belong to “Not anaemic”.

The last model correctly classified 1622 (76.8%) women as “Any anaemic” and 4091 (89.2%) women as “Not anaemic”. The remaining 490 (23.2%) and 494 (10.8%) women were classified incorrectly as “Not anaemic” and “Any anaemic” while in fact they belong to “Any anaemic” and “Not anaemic”, respectively.

From these results, it is possible to say that the J48 unpruned model with all attributes is the best model in terms of minimized incorrectly classified instances.

5.1.2 Modeling using PART rule induction

The second data mining technique used in this study was PART rule induction classifier. To build this model, like J48 decision tree, two experiments were conducted and four scenarios were considered, two containing all 22 attributes and the other two containing selected 13 attributes. The results obtained from these experiments were summarized in Table 5.5 with their respective performance measures.

Table 5.5: PART rule induction experiment results – Final EDHS 2011 Dataset

Experiment	Model	Detailed Performance Measures							
		Accuracy	TP Rate	TN Rate	Precision	Recall	F-Measures	ROC area	Speed
1	PART Pruned with all attributes	93.6%	90.5%	95%	0.893	0.905	0.899	0.973	2.08
	PART Pruned with selected attributes	84.8%	75.9%	88.9%	0.76	0.759	0.76	0.918	1.13
2	PART Unpruned with all attributes	95.2%	93%	96.2%	0.918	0.93	0.924	0.977	2.52
	PART Unpruned with selected attributes	85.8%	78.7%	89%	0.767	0.787	0.777	0.936	1.62

As it is shown in Table 5.5, two experiments were conducted. The first experiment was designed to evaluate the performance of PART pruned model in predicting any anaemic women aged 15-49 years, whereas the second experiment was designed to evaluate the performance of PART unpruned classifier. In each experiment, two scenarios were considered and the first scenario of each experiment contained all 22 attributes and the second scenario contained selected 13 attributes.

In the first scenario of each experiment the model was developed on a full training set containing 6,697 instances with all 22 attributes. In the first experiment scenario one and two took 2.08 and 1.13 seconds respectively to build the model, while in the second experiment, scenario one and two took 2.52 and 1.62 seconds.

While modelling using PART rule induction algorithm, different results were obtained during the two experiments. In the first experiment, the result of PART pruned model with all attributes shows that the model correctly classified 6267 (93.6%) instances, while 430 (6.4%) of the instances were classified incorrectly. The result of PART pruned with selected attributes also indicates that the model correctly classified 5682 (84.8%) instances while 1015 (15.2%) of the instances were classified incorrectly.

The second experiment indicates that the model built using PART unpruned classifier with all attributes correctly classified 6374 (95.2%) instances while 323 (4.8%) of the instances were classified incorrectly. The result of PART unpruned with selected attributes also shows that the model correctly classified 5744 (85.8%) instances while 953 (14.2%) of the instances were classified incorrectly.

Considering these results, the model performance with the pruned and unpruned scenarios when using all attributes perform better than when using selected attributes with their respective performance measures, except the speed. However, in modelling using PART rule induction technique, the PART unpruned tree model with all attributes performs better compared to other models and as a result it is possible to say it is the best model for prediction. The experimental outputs of the PART unpruned model with all attributes is presented in ANNEX 4.

Regarding the confusion matrix of the PART rule induction technique, the result summaries are shown in Tables 5.6 and 5.7, respectively.

Table 5.6: Confusion Matrix of PART rule induction for Experiment 1

Model 1	**Confusion Matrix**	
PART pruned with all attributes	Any anaemic	Not anaemic
	1912(90.5%)	200 (9.5%)
	230 (5.1%)	4355 (94.9%)

Model 2	Confusion Matrix	
PART pruned with selected attributes	Any anaemic	Not anaemic
	1604 (76%)	508 (24%)
	507 (11%)	4078 (89%)

As it is indicated in Table 5.6, model one correctly classified 1912 (90.5%) woman out of 2112 women as "Any anaemic". That means, 200 (9.5%) women were incorrectly classified as "Not anaemic", while in fact they belong to "Any anaemic". Results in model one also depict that 4355 (94.9%) women were correctly classified as "Not anaemic" and the remaining 230 (5.1%) women were incorrectly classified as "Any anaemic", while they actually belong to "Not anaemic".

The second model also correctly classified 1604 (76%) women as "Any anaemic" and, on the other hand, 508 (24%) women misclassified as "Not anaemic", while they should be "Any anaemic". This model also correctly classified 4078 (89%) women as "Not anaemic" and the remaining 507 (11%) women were misclassified as "Any anaemic", while in fact they belong to "Not anaemic".

Table 5.7: Confusion Matrix of PART rule induction for Experiment 2

Model 3	Confusion Matrix	
PART Unpruned with all attributes	Any anaemic	Not anaemic
	1964 (93%)	148 (7%)
	175 (4%)	4410 (96%)
Model 4	**Confusion Matrix**	
PART Unpruned with selected attributes	Any anaemic	Not anaemic
	1663(78.7%)	449 (21.3%)
	504 (10.9%)	4081 (89.1%)

The results in Table 5.7 show that, model three correctly classified 1964 (93%) women out of 2112 as "Any anaemic" and the remaining 148 (7%) women were classified incorrectly as

"Not anaemic", while in fact they belong to "Any anaemic". The model also correctly classified 4410 (96%) women as "Not anaemic" and misclassified 175 (4%) women as "Any anaemic", while they actually belong to "Not anaemic".

The fourth model correctly classified 1663 (78.7%) women as "Any anaemic" and 4081 (89.1%) women as "Not anaemic". The remaining 449 (21.3%) and 504 (10.9%) women were misclassified as "Not anaemic" and "Any anaemic", while they actually belong to "Any anaemic" and "Not anaemic", respectively.

From these results, it is also possible to say that the PART unpruned model with all attributes is considered as the best model in terms of minimized incorrectly classified instances.

5.2 Performance evaluation of classifiers

After building the models by conducting different experiments using two categories, the next step was comparing and evaluating the performance of the models so as to select the best model which is capable of predicting the status of anaemia in women of reproductive age.

To compare and evaluate the performance of the models, the standard performance metrics like accuracy, TP Rate, TN Rate, Precision, Recall, and F-Measure were used. ROC area and speed were also other parameters used to compare model's performance. Generally, the comparison was performed based on the results of the aforementioned standard performance evaluation measures provided for the models used in the experiments.

As shown above in Table 5.2 and Table 5.5, the values varied between those eight models. For instance, the accuracy of the models varies between 84.3 - 95.2 percent. It includes lower scores for J48 pruned model with selected attributes and higher scores for PART unpruned model with all attributes. The ROC area value also range between 0.907 – 0.977. It includes lowest scores for J48 pruned with selected attributes and higher scores for both J48 and PART unpruned models with all attributes.

From the precision values, the highest scores of 0.918 were registered by PART Unpruned with all attributes followed by J48 unpruned model with all attributes that achieved 0.91. The least scores of 0.714 were registered by the PART unpruned model with selected attributes. When we come to the Recall and F-Measure results, PART unpruned model with all attributes achieved the highest scores of 0.93 and 0.924 respectively, whereas the least Recall

rate of 0.715 and F-Measure of 0.742 were presented by the J48 pruned model with selected attributes.

The other performance measures used to compare the models were TP Rate (Sensitivity) and TN Rate (Specificity). From the result the highest TP Rate of 93% was scored by PART unpruned model with all attributes, while the least TP Rate of 71.5% was scored by J48 pruned with selected attributes. The highest TN Rate of 96.2% was registered by PART unpruned model with all attributes, whereas the lowest TN Rate of 88.9% was obtained by PART pruned with selected attributes. Here, a quick look into the result revealed the TN Rate (Specificity) scores of the models are slightly higher than the TP Rate (Sensitivity) scores of the models. This means, all models were better in predicting negative cases compared to positive cases.

Regarding the speed, i.e. the total training time in seconds, the J48 model with selected attributes took less time followed by the model with all attributes. Here, the result indicates that, having reduced number of attributes has a benefit of fast execution time. In general, based on the time taken to build a model, the J48 decision tree techniques with selected attributes and all input attributes achieved the quickest execution time than the PART rule induction model.

From the above performance results, we can observe that the models built using all input attributes achieved the best performance compared to others models built using selected attributes. In other words, it has been found that the models using selected attributes achieved lowest performance than using all attributes. Ideally, the use of reduced or selected attributes in building a model assumed to optimize the correctness accuracy and also increases the overall performance of the model. However, the result of this study as indicated above shows the opposite. This result was not surprising as Biset (38) and Tesfahun (45) compared the effect of reduced attributes on the model performance and found that the model built with all input attributes outperformed that of the model with selected attributes for most of the standard performance metrics.

Generally, the analysis of the experimental result shows that, the PART unpruned model with all attributes outperformed the other models by achieving the highest Accuracy, TP Rate, TN Rate, Precision, Recall, F-Measure. Both J48 and PART unpruned model with all attributes

obtained equal scores of ROC area, whereas the J48 models with selected and all attributes achieved the fastest execution time.

Therefore, based on these performance results, the PART unpruned model with all attributes was selected as the best model for this study. This model has achieved an outstanding performance of 95.2% accuracy.

Although the predictive performance accuracy of the selected model was very good, i.e. 95.2%, the model still commits an accuracy error. That means, the model incorrectly classified 4.8% of the instances to some other classes. The possible reason may be due to the nature of the dataset used in this study.

5.3 Rule Extraction

Both classifiers such as J48 Decision Tree and PART Rule Induction used in this study produce important and interesting rules. In this part, however, the rules extracted using the PART unpruned model with all attributes, which was selected as the best model, were presented. Among the corresponding rules extracted by this model, some of the rules that cover most of the instances in the dataset are selected and presented along with their interpretations which make each rule understandable by the users as follows:

Rule 1: IF Pregnancy Status = Yes AND Body Mass Index = Low AND AND Age = 30-34AND Wealth =Poor AND Iron Supplementation = No THEN Any anaemic.

This rule is interpreted as: a woman who has pregnancy and whose age is between 30 to 34, and her body mass index is low, and poor in wealth, and not taken iron, there is high probability of any anaemic disease for the women.

Rule 2: IF Residence= Rural AND Educational Level = No education AND Region =Afar AND Visited Health Facility = No AND Age = 25-29 THEN Any anaemic.

This is also interpreted as: women who lived in Afar region of rural area and whose age is between 25 to 29, and if not educated, and not visited the health facility, there is high probability of any anaemic disease for the women.

Rule 3: IF Visited Health Facility = No AND Educational Level = No education AND Contraceptive use = No AND Age = 25-29 THEN Any anaemic.

When interpreted, it is stated as: women who is not visiting health facility and whose age is between 25 to 29, and if not educated, and not used contraceptive method, the probability of any anaemic women is high.

Rule 4: IF Body Mass Index = Normal AND Pregnancy Status = Yes AND Antenatal Care = No AND Age = 35-39 THEN Any Anaemic.

Rule four is interpreted as: women whose weight is normal and age is between 35 to 39, and if she is pregnant, and not followed the antenatal care, the probability of any anaemic women is high.

Rule 5: IF Educational Level = No education AND Region = Somali AND Residence = Rural AND Marital Status = Never married AND Age = 20-24 THEN Any anaemic.

Finally, rule five is interpreted as: women who is not educated and whose age is between 20 to 24, and who lived in Afar region of rural area, and never married, then any anaemic women will be expected.

As it is observed from rules 1 to 5 above, any anaemia is significantly associated with Educational Level, Contraceptive use, Residence, Wealth, Age, Pregnancy Status, Visited Health Facility, Region, Marital Status, Chewed chat, Body Mass Index, and Alcohol Intake. It is also indicated in the rules that the age category of women between 25-29, 30-34, and 35-39 years were found as the most vulnerable groups to anaemia due to the fact that the age category is fertility intensive in women's life.

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

5.4 The Potential Set of Attributes

One of the objectives of this study was to discover the potential set of attributes which can reasonably predict any anaemic women. In this research, 22 attributes were used for building a model. To discover the potential set of attributes out of those 22 attributes, the researcher used the performance results and the results of the rule extracted from the PART unpruned model with all attributes.

On the basis of such results, the following set of attributes as indicated in Table 5.8 were reckoned as potentially the most determinant factors to predict any anaemic women.

Table 5.8: The Potential set of attributes list – EDHS 2011 Dataset

No.	Attribute Name	Descriptions
1	AGE	Women's age in years
2	RESIDENCE	Type of place of residence
3	EDUCLEV	Highest educational level
4	WEALTH	Respondent's wealth index
5	BMI	Body mass index of person
6	IRON	Iron supplementation for women of the reproductive age
7	PREGNANCY	Current pregnancy status
8	CONTRACEPTIVE	Current contraceptive use
9	HEALTHF	Visited health facility in last 12 months
10	ANC	Antenatal care received by women
11	REGION	Region where the survey conducted
12	MARRY	Current marital status

To confirm these attributes as potentially determinant factors, the researcher conducted experiments using the discovered 12 potential set of attributes. The results obtained from these experiments were summarized in Table 5.9 with their respective performance measures.

Table 5.9: Results of experiments with 12 potential set of attributes

Experiment	Model	Performance Measures	
		Accuracy	**ROC area**
1	J48 Pruned with potential set of attributes	87.3%	0.933
2	J48 Unpruned with potential set of attributes	89.4%	0.95
3	PART Pruned with potential set of attributes	88.4%	0.943
4	PART Unpruned with potential set of attributes	89.8%	0.954

As we can see from Table 5.9, the result shows that the models with those 12 potential set of attributes generally makes more than 3 % improvement on the models accuracy and an increase on the value of ROC area compared with selected 13 attributes as indicated in the above Table 5.2 and Table 5.5. This ensures that, the identified attributes are potentially the determinant risk factors for anaemia.

Having those results with potential set of attributes then, the researcher consulted domain experts particularly the nutritionist, and some previous studies for other confirmation. Although the domain experts believed that the remaining attributes had also a role in prediction, they agreed on such attributes which were identified as potential factors of anaemia. Moreover, some previous studies, though reported on various patterns, also ensure these attributes as determinant risk factors for anaemia in women of reproductive age. In this regard, the EDHS 2011 (8) documented pregnancy, rural residence, some regions, no education, and wealth status as determinant factors for the prevalence of anaemia in women. A study by Samson and Fikre (15) reported rural residence, poor educational and economic status; 25-39 years of age, high parity and pregnancy were key factors. Similar studies by Sanku (35) and Bentley (36) also reported that pregnant women, poorest women, women of age 20-24 years, women not taking iron, not using contraceptive, and visiting health centres, women with habit of cigrates smoking and others had a relatively high prevalence of anaemia.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

In this chapter, we conclude the study in four parts: summary of the thesis, answers to the research objectives, strength and limitation of the study, and finally recommendations.

6.1 Summary of the Thesis

This thesis is based on a quantitative study designed to apply data mining techniques on the secondary datasets, the EDHS 2011, which was designed to provide representative estimates of health and demographic indicators at national level and across regions. The target attributes for this study were either any anaemic or not anaemic women aged 15- 49 years. The study implemented the KDD methodology. Initially, 25 attributes out of 920 variables were selected to build the final dataset. By finishing all preparation efforts, the attribute counts had been reduced to 22 from its original 25, by using personal judgment of the researcher based on relevance to the targets, medical expert, and results of previous studies. The final study dataset included 7,2,9 and 4 input attributes from Socio-demographic, Nutritional status, Reproductive health related and Health behaviour, respectively.

To build the models using the supervised data mining techniques, mainly J48 decision tree and PART rule induction, four experiments using two categories, one containing all 22 attributes and the second containing selected 13 attributes were carried out and as a result a total of eight models were developed. From the performance results, it has been found that the models built using all attributes well performed with a remarkable predictive result in all of those performance measures compared to models built using selected attributes.

Among the models built, the PART unpruned model using all attributes scored the highest correctness accuracy of 95.2%, while J48 pruned with selected attributes scored the least accuracy of 84.3%. However, both PART and J48 unpruned models with all attributes returned equal ROC area value of 0.977, which was the highest score. Based on the analysis of the performance result, the PART unpruned model with all attributes outperformed the other models by achieving the highest Accuracy, TP Rate, TN Rate, Precision, Recall, F-Measure, and ROC area. Higher scores in almost all of the standard performance measures which were implemented in this study has signified the capability of PART unpruned model with all attributes in predicting the status of anaemia in women of reproductive age.

Moreover, this model also obtained the lower error rate compared to other models. Thus, the model, PART unpruned with all attributes is selected as the best model in the study.

Rules extracted using the selected models were used to reach to the potential set of attributes responsible for prediction. From the rules, some interesting rules that include attributes believed to be as determinant risk factors for anaemia were found. Having the identified attributes using those rules, and performance results and getting a confirmation from domain experts and some previous studies, the study finally introduced the potential set of attributes which can reasonably predict any anaemia.

The findings clearly suggest that those discovered potential set of attributes have strong association with prevalence of anaemia in women of reproductive age. Therefore, the results from this study can help health planners to design efficient surveys at local, regional or national levels, and track the major determinant factors for anaemia among women of reproductive age. And they can use their customized survey to better plan towards reducing anaemia prevalence and other health initiatives.

6.2 Research objectives

The study has successfully achieved all its three objectives, which included building predictive models using decision tree and rule induction algorithm, comparing the models based on standard performance measures, and discovering the potential set of attributes which can reasonably predict anaemia in women.

With respect to the first objective, we conclude that various predictive models, in this case, eight models were built to estimate any anaemic women.

Regarding the second objective, the researcher compared and evaluated their performance by combination of standard metrics such as Accuracy, TP Rate, TN Rate, Precision, Recall, F-Measure, ROC area, and speed. From the result, it could be concluded that the PART unpruned model with all attributes had the highest scores in all of such performance measures.

With respect to the third objective, after discovering and confirmation, the 12 attributes indicated in Table 5.9 are recommended as the potential set of attributes to predict any anaemic women of reproductive age.

6.3 Strength and Limitation of the Study

With respect to the strength of the study, although the intended classification accuracy was not achieved, with very good predictive performance accuracy, the study result ensures the potential capability of data mining techniques application in the field of health care. Further, since the objectives of this research were achieved, we can conclude that, overall, the study was successful.

Basically, in data mining techniques, the intention was to build a predictive model with an accuracy of 100%. However, in this study due to the nature of the 2011 EDHS dataset itself limits the predictive model did not achieve the intended classification accuracy. Another limitation of this study was lack of prior experience to data mining technology as well as to the problem domain was also the limitations faced by the researcher.

The major limitation of this research is, due to time limitation the researcher is unable to investigate further for some results of the study like the outperformance of models with all attributes compared to selected attributes, and the less execution time of J48 decision tree models compared to PART rule induction models.

6.4 Recommendations

Based on the findings of the research, the following recommendations are made:

- Health planners together with other stakeholders should use the proposed potential set of attributes to design a proper and suitable preventive and disease-control programs to combat anaemia.
- Policy makers and health service providers should use and implement the findings to real field practices and provisions.
- It is also recommended that the identified model should be applied in the field and monitored to validate the finding.
- Finally, as this study was conducted on one of the vulnerable group of anaemia, i.e. women aged 15-49 years, it is necessary to carry out on other groups like preschool and school-age children, and the elderly to investigate the prevalence of anaemia at national level.

REFERENCES

1. World Health Organization. The Prevalence of anaemia in women: a tabulation of available information, 2nd ed., 1992.
2. Worldwide prevalence of anaemia 1993–2005: WHO global database on anaemia Geneva: WHO. 2008. [Cited 2012 Nov 23]. Available from: *http://whqlibdoc.who.int/publications/2008/9789241596657_eng.pdf*
3. Sharman A. Anaemia testing in population-based surveys: general information and guidelines for country monitors and program managers. Calverton, Maryland USA: ORC Macro. 2000.
4. Royston E. The prevalence of nutritional anaemia in women in developing countries: a critical review of available information. *World Health Stat Q 1982;35:52–91*.
5. UNICEF/UNU/WHO. Iron deficiency anaemia: assessment, prevention, and control. A guide for programme managers. WHO/NHD, 2001 [report no. 01.3].
6. World Health Organization. The world health report 2000 – health systems: improving performance. Geneva: WHO, 2000.
7. World Health Organization. The World health report 2002: reducing risk and promoting healthy life. The institute; [Cited 2012 Nov 20]. Available from: *http://www.who.int/whr/2002/en/*.
8. Central Statistical Agency [Ethiopia] and ICF International. Ethiopia Demographic and Health Survey 2011. Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International. 2012.
9. ORC Macro and CSA. Ethiopian Demographic and Health Survey 2005, 2006.
10. United Nations Development Program Emergency Unit for Ethiopia. Primary health Care in Ethiopia . University of Pennsylvania; 2000 [Cited 2012 Nov 11]. Available from: *http://www.africa.upenn.edu/eue_web/eue_mnu.htm*.
11. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. New York: John Wiley & Sons, 2003.
12. David H, Heikki M, Padhraic S. Principles of Data Mining. The MIT Press, 2001.
13. Daniel T. L. Discovering knowledge in data: An introduction to data mining. Canada: John Wiley & sons, 2005.
14. Han J, Kamber, M. Data mining: concepts and techniques, 2nd ed. San Fransisco, CA: Morgan Kaufmann, 2006.

15. Samson G, Fikre E. Correlates of anemia among women of reproductive age in Ethiopia: Evidence from Ethiopian DHS 2005. *Ethiop. J. Health Dev. 2011;25(1):22-30.*
16. Haidar J. Prevalence of Anaemia, Deficiencies of Iron and Folic Acid and Their Determinants in Ethiopian Women. *J Health Popul Nutr. 2010: 28(4): 359–368.*
17. Haidar J, Pobocik R. Iron Deficiency Anemia is not a rare problem among women of reproductive ages in Ethiopia: A Community based cross-sectional study. *BMC Blood Disorders 2009:9(7).*
18. Central Statistical Agency (CSA). [Ethiopia]. The 2007 Population and Housing Census of Ethiopia. Statistical Summary Report at National Level. Addis Ababa, Ethiopia: Central Statistical Agency. 2008.
19. Thearling, K. An introduction to data mining: discovering hidden value in your data warehouse. [Internet]. [Cited 2013 January 08]. Available from: www.thearling.com/text/dmwhite.htm
20. Berson, A., Smith, S., and Thearling, K. An overview of data mining techniques. In: Building data mining applications for CRM. [Internet]. [Cited 2013 January 08]. Available from: www.thearling.com/text/dmtechniques/dmtechniques.htm
21. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. Third Edition. Potomac, USA: Two Crows Corporation, 2005.
22. Moxon, B. "Defining data mining, the how's and whys of data mining, and it differs from other analytical techniques". Online addition of DBMS data warehouse supplement. 1996.
23. Groth, R. Data mining: a hand on approach for business professionals. New Jersey: Prentice-Hall, 1998.
24. Witten, I. H. and Frank, E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, CA, Morgan Kaufman, 2005.
25. Milley, A. Healthcare and data mining. *Health Management Technology*, 2000. 21(8), 44-47.
26. Eapen, A. G. Application of Data mining in Medical Applications. Ontario, Canada: University of Waterloo, 2004.
27. Friedman, J. H. *Data Mining and Statistics: What's the Connection?* [Cited 2013 January 02]. Available from: http://www-stat.stanford.edu/~jhf/ftp/dm-stat.pdf

28. Johnson, R., Wicheren,D.W. Applied Multivariate Statistical Analysis. New York: Prentice-Hall, 1998.

29. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Eds. Advances in Knowledge Discovery and Data Mining. Cambridge, Mass. : AAAI/MIT Press, 1996.

30. Chapman, P. et al,. CRISP-DM 1.0 - Step-by-step data mining guide. [Internet]. 2000. [Cited 2013 January 10]. Available from: http://www.crisp-dm.org/CRISPWP-0800.pdf

31. SAS Enterprise Miner – SEMMA. [Cited 2013 January 10]. Available from: http://www.sas.com/technologies/analytics/datamining/miner/semma.html.

32. Langley, P., Simon, H. A. "Application of Machine Learning and Rule Induction". 1995, *Communications of the ACM*, (38:11), 55.64.

33. Maimon, oded and Rokach, Lior, eds. Data mining and knowledge discovery handbook. 2nd ed. London: Springer, 2010.

34. Weiss, Sholom M. and Zhang, Tong. Performance analysis and evaluation. In:Ye Nong, editor. The Hand book of data mining. New Jeresy, USA: Lawerence Erlbaum Associates Inc., 2003.

35. Sanku D, Sankar G, and Madhuvhhanda G. Prevalence of anaemia in women of reproductive age in Meghalaya: a logistic regression analysis. Turk J Medical Sci. 2010, 40(5): 783-789. [Cited 2012 December 15]. Available from: journals.tubitak.gov.tr/medical/issues/sag-1040-5/sag-40-5-17-0811-44.pdf

36. Bentley, M and Griffiths, P. The burden of anaemia among women in India. European Journal of Clinical Nutrition. 2003. 57, 52-60. [Cited 2012 December 15]. Available from: www.idpas.org/pdf/1885TheBurdenofAnemiaAmongWomenInIndia.pdf

37. Sanap, S, Nagori, M, and Kshirsager, V. Classification of anaemia using data mining techniques. 2011. Swarm, Revolutionary, and Memetic Computing. Lecture notes in computer science, volume 7077, pp 113-121. [Cited 2013 January 05]. Available from: link.springer.com/chapter/10.1007/1978-3-642-27242-4-14?no-access=true

38. Biset D. Predicting low birth weight using data mining techniques on Ethiopian Demographic and Health Survey data sets. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2011.

39. Abel Damtew. Designing a predictive model for heart disease detection using data mining techniques. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2011.

40. Muluneh Endalew. Exploring the prevalence of diarrheal disease using data mining technology: a case of Tikur Anbessa Hospital. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2011.

41. Elias Lemuye. HIV status predictive modelling using data mining technology. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2011.

42. Selam Assamnew. Predicting the occurrence of measles outbreak in Ethiopia using data mining technology. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2011.

43. Shegaw A. Application of data mining technology to predict child mortality patterns: the case of Butajira Rural Health Project (BRHP). [M.sc. thesis]. Addis Ababa University, Ethiopia. 2002.

44. Teklu U. Application of data mining techniques on Antiretroviral Therapy (ART) data: the case of Adama and Asella hospitals. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2010.

45. Tesfahun Hailemariam. Application of data mining for predicting adult mortality. [M.sc. thesis]. Addis Ababa University, Ethiopia. 2012.

46. World Health Organization (WHO). Nutritional anaemias. Report of a WHO Scientific Group. Geneva, Switzerland: WHO, 1968.

47. Fleming, A. F. Maternal anemia in northern Nigeria: Causes and solutions. 1987. *World Health Forum* 8:339-343.

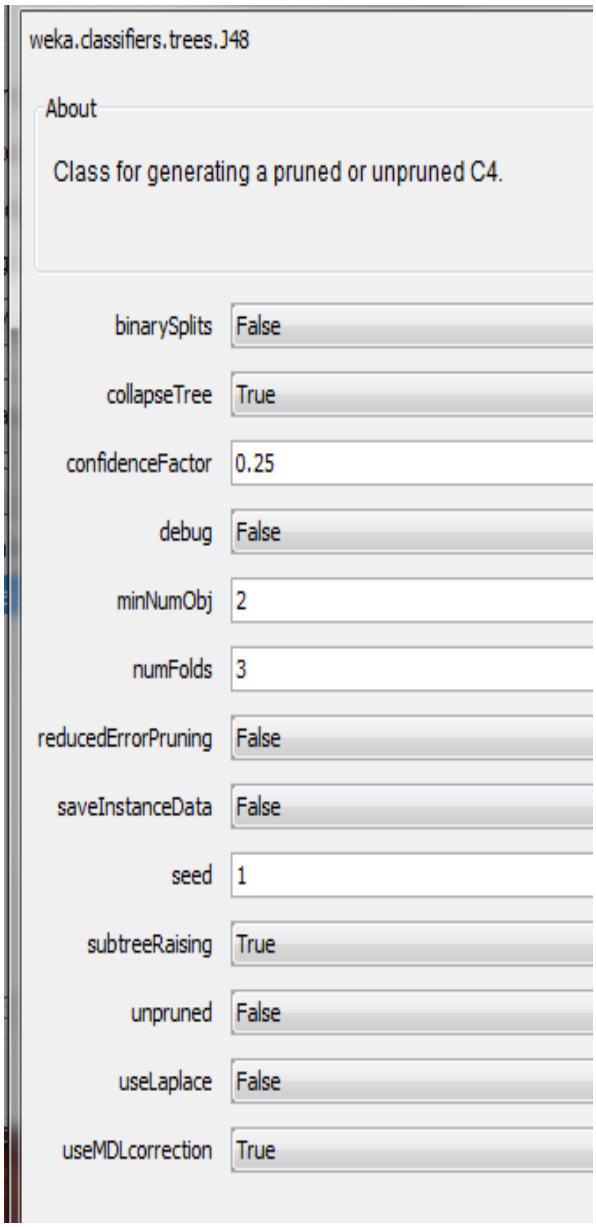
48. Berry J.A.M, Linoff G. Data mining techniques for marketing sales and customer support. New York: Wiley, 2004.

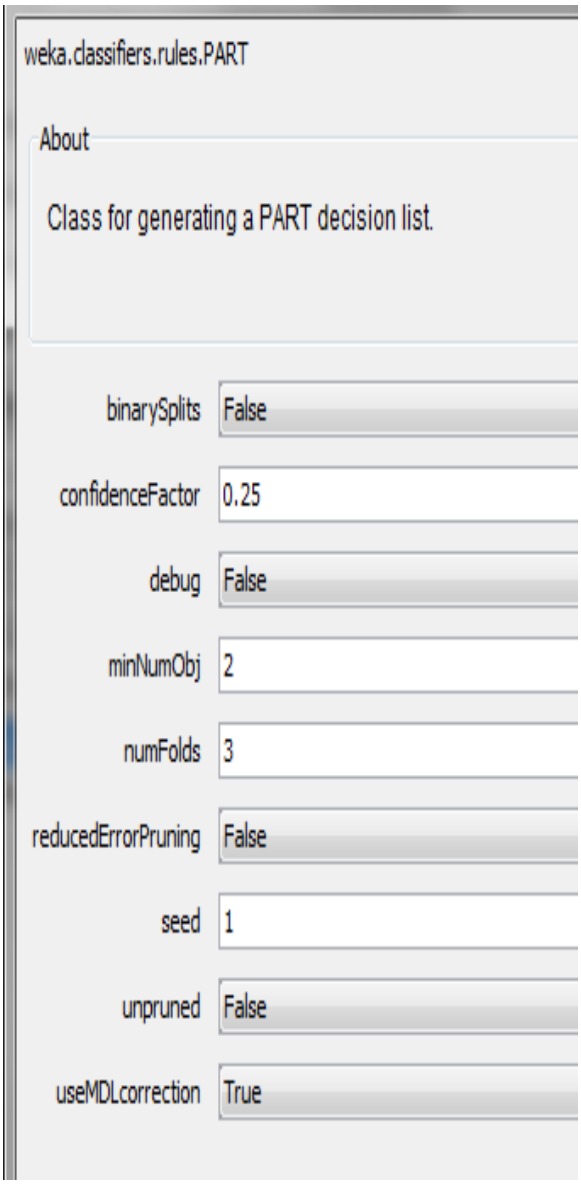
49. Kim, Y. Street, and Menczer, F. Feature selection in Data Mining. 2000. [Cited 2013 April 15]. University of Iowa, USA. Available from: http://dollar.biz.uiowa.edu/~street/research/dmoc.pdf .

ANNEX 1: List of Consulted Domain Experts

No.	Name	Specialization	Institute
1	Dr. Jemal Haider	Nutritionist	School of Public Health, A.A.U
2	Dr. Solomon Shiferaw	Nutritionist	School of Public Health, A.A.U

ANNEX 2: Parameter settings of the two techniques used in conducting the experiments





ANNEX 3: Run information of J48 Unpruned decision tree model with all attributes

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

Relation: Final List of 22 Attributes

Instances: 6697

Attributes: 23

Age

Region

Residence

Educational Level

Wealth

Give Birth

Pregnancy Status

Terminated pregnancy

Contraceptive use

Visited health facility

Breastfeeding status

BodyMassIndex

Smoke cigarets

Smokes other

Marital Status

Employment

Place of delivery

Postnatal checkup

Antenatal Care

Iron supplement

Chewed chat

Alcholic Intake

Anaemia Level

Test mode: 10-fold cross-validation

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

J48 unpruned tree

Number of Leaves : 899

Size of the tree : 1232

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	6333	94.5647 %
Incorrectly Classified Instances	364	5.4353 %
Mean absolute error	0.064	
Root mean squared error	0.2057	
Relative absolute error	14.8232 %	
Root relative squared error	44.2648 %	
Coverage of cases (0.95 level)	98.4172 %	
Mean rel. region size (0.95 level)	56.4133 %	
Total Number of Instances	6697	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.919	0.042	0.91	0.919	0.914	0.875	0.977	0.948	Any anaemic
	0.958	0.081	0.963	0.958	0.96	0.875	0.977	0.984	Not anaemic
Weighted Avg.	0.946	0.069	0.946	0.946	0.946	0.875	0.977	0.973	

=== Confusion Matrix ===

a	b	<-- classified as
1941	171	a = Any anaemic
193	4392	b = Not anaemic

ANNEX 4: Run information of PART Unpruned rule induction model with all attributes

Scheme: weka.classifiers.rules.PART -U -M 2 -C 0.25 -Q 1

Relation: Final List of 22 Attributes

Instances: 6697

Attributes: 23

Age

Region

Residence

Educational Level

Wealth

Give Birth

Pregnancy Status

Terminated pregnancy

Contraceptive use

Visited health facility

Breastfeeding status

BodyMassIndex

Smoke cigarets

Smokes other

Marital Status

Employment

Place of delivery

Postnatal checkup

Antenatal Care

Iron supplement

Chewed chat

Alcholic Intake

Anaemia Level

Test mode: 10-fold cross-validation

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

PART decision list

Number of Rules : 380

Time taken to build model: 2.52 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	6374	95.1769 %
Incorrectly Classified Instances	323	4.8231 %
Kappa statistic	0.8887	
Mean absolute error	0.0505	
Root mean squared error	0.1878	
Relative absolute error	11.704 %	
Root relative squared error	40.4266 %	
Coverage of cases (0.95 level)	98.2977 %	
Mean rel. region size (0.95 level)	53.7629 %	
Total Number of Instances	6697	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.93	0.038	0.918	0.93	0.924	0.889	0.977	0.946	Any anaemic
	0.962	0.07	0.968	0.962	0.965	0.889	0.977	0.982	Not anaemic
Weighted Avg.	0.952	0.06	0.952	0.952	0.952	0.889	0.977	0.971	

=== Confusion Matrix ===

a	b	<-- classified as
1964	148	a = Any anaemic
175	4410	b = Not anaemic