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

PREDICTING MATERNAL HEALTH CARE SEEKING PATTERN USING DATA MINING TECHNIQUES IN ETHIOPIA

DAWIT AYELE

JUNE 2013

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

PREDICTING MATERNAL HEALTH CARE SEEKING PATTERN USING DATA MINING TECHNIQUES IN ETHIOPIA

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

BY

DAWIT AYELE

JUNE 2013

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

PREDICTING MATERNAL HEALTH CARE SEEKING PATTERN USING DATA MINING TECHNIQUES IN ETHIOPIA

BY

DAWIT AYELE

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

_____

Dawit Ayele

JUNE, 2013

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

_____

Ato Getachew Jemaneh

JUNE, 2013

_____

Wubegzier Mekonnen (PhD)

JUNE, 2013

DEDICATION

I dedicate this study to my family who have been supporting and caring for my education.

ACKNOWLEDGMENTS

First of all, I would like to thank the almighty GOD for giving me the ability to pursue the study.

My honorable gratitude extends to my advisors Ato Getachew Jemaneh and Dr. Wubegzier Mekonnen for the continuous support and guidance in the whole undertakings associated with my study. Had it not been their diligence and commitment, this tough endeavor would have ended up in vain.

Addis Ababa University School of graduate studies department of Health Informatics is the other stakeholder that deserve the due acknowledgement for its logistics and financial support geared for the realization of the study.

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

A great deal of thanks should also be granted to my friends Misganaw Tadesse, Teketel Mulugeta, Zenebe Markos and Addisalem Abebe with whom I shared a lot of experiences pertaining to the study.

Last but not least, I owe my heartfelt gratitude to my family for their passion and sympathy that become a source of inspiration for me to effectively conduct the research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ANC	Antenatal Care
CRISP	Cross Industry Standard Process
CSAE	Central Statistics Agency of Ethiopia
CSV	Comma Separated Value
DM	Data Mining
EDHS	Ethiopian Demographic and Health Survey
FMOH	Federal Ministry of Health
HIV	Human Immunodeficiency Virus
HSDP	Health Development Sector Development Program
ICF	International Confederation of Trade
KDP	Knowledge Discovery Process
KDD	Knowledge Discovery in Databases
MHCS	Maternal Health Care Seeking
MDG	Millennium Development Goal
MMR	Maternal Mortality Ratio
NGOS	Non-Governmental Organizations
ROC	Receiver Operating Characteristics
SAS	Statistical Analysis Software
SPSS	Statistical Package for Social Sciences
STI	Sexually Transmitted Infection
UNFPA	United Nations Population Fund
UNICEF	United Nations International Children's Fund
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSTRACT

Background: Utilization of maternal health care services could save unnecessary severe complications and death among women during pregnancy, delivery and after delivery. Ethiopia is one of the countries with high maternal morbidity and mortality in sub-Saharan Africa which needs more public health care effort in the country.

Objective: The general objective of the study was to construct a model that can predict the maternal health care seeking pattern of reproductive age in Ethiopia.

Methodology: The study followed Hybrid methodology of Knowledge Discovery Process to achieve the goal of building predictive model using data mining techniques.

Therefore, the overall research design was to build a model that can predict the maternal health care seeking pattern using J48 Decision tree and Naïve Bayes algorithms in Ethiopia from EDHS 2011 dataset. WEKA 3.6.8 data mining tools and techniques were employed as a means to address the research problem.

Results: The result of the study showed that the J48 Decision tree algorithm outperforms Naïve Bayes on the three of the outcome variables. For antenatal care the model was selected with an accuracy of 74.78%. Then for the second outcome variable (delivery care) the model was selected with an accuracy of 91.23% and area under receiver operating characteristics of 0.89. Finally for postnatal care the model was selected with an accuracy of 87.5% and area under receiver operating characteristics curve of 0.80. The best attributes selected for each of the outcome variables are Place of Residence, Household Wealth Index, Women's Educational level, Husbands Occupation, Region, Husbands Educational level, Total number of children, Media Exposure.

Conclusion: In general, the results obtained from this study were interesting and encouraging; it can be used as decision support for healthcare practitioner. The finding shows that there is a regional difference in utilizing maternal health care service in the country, thus it is recommended that all the concerned parties should give due consideration for these regions, increasing maternal education at least up to primary level in all regions of the country, provision of opportunities for employment and poverty reduction especially in rural parts of the region.

CHAPTER ONE

INTRODUCTION

1.1 Background

1.1.1 Maternal Health in a Global Context

According to WHO maternal death is defined as the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes [1].

Improving maternal health is one of the eight Millennium Development Goals (MDGs) [2]. According to estimates developed by the WHO, UNICEF, UNFPA and the World Bank, there were 358,000 maternal deaths globally during 2008 [3]. Of the total estimated maternal deaths, developing countries accounted for 99% maternal deaths and with 63,000 cases in the year 2008.

The two targets for assessing MDG 5 are reducing the maternal mortality ratio (MMR) by three quarters between 1990 and 2015, and achieving universal access to reproductive health by 2015.

Worldwide over half a million women die as a result of childbirth or complication due to pregnancy. Almost all (99%) of these deaths occur in developing countries. Asia and Africa alone take 95% of the share of the world's maternal death and the number is almost equally divided between Asia (253,000) and Africa (251,000). Four percent of the deaths occur in Latin America and the remaining one percent in the more developed regions of the world [4].

1.1.2 Maternal Health in Developing Countries

The child bearing functions of women, especially in developing countries, have been granted as a normal or routine process. Yet these valued and precious parts of life are among the most hazardous experiences that women often engage in without being aware of the risks or dangers that they are in [5].

The WHO estimates that 580,000 women of reproductive age die each year from complications arising from pregnancy, and a high proportion of these deaths occur in sub-Saharan Africa [5]. The ratio of maternal mortality in the region is one of the highest in the world, reaching levels of 686 per 100,000 live births in 1994 [6]. Women play a principal

role in the rearing of children and the management of family affairs, and their loss from maternity-related causes is a significant social and personal tragedy [6].

Studies demonstrating the high levels of maternal mortality and morbidity in developing countries and research identifying causes of maternal deaths have repeatedly emphasized the need for antenatal care and availability of trained personnel to attend women during labor and delivery [6, 7, 8]. The importance of tetanus toxoid injections given prior to birth to reduce neonatal mortality has been documented as well [9].

1.1.3 Maternal Health in Ethiopia

Maternal health status in Ethiopia is one of the worst in the world. The country is characterized by high maternal and child mortality. The maternal mortality rate was estimated at 673/100,000 according to the 2005 EDHS and infant mortality rate was 77/1000 [10]. It is noted that there has been a minimal change in maternal mortality in five years from 871/100,000 in 2000 to 673/100,000 in 2005 [10, 11]. According to the 2011 EDHS report, the maternal mortality ratio is 676 deaths per 100,000 live births. The estimated maternal mortality ratio is almost the same in the 2011 EDHS (676) as it was in the 2005 EDHS (673).

The observed change in maternal mortality is very low and there is a need to accelerate the decline in mortality in order to achieve the MDG of reducing maternal mortality by two thirds. Among the leading causes of maternal death in the country haemorrhage constitutes (25%) followed by puerperal infections (15%), eclampsia (13%) and complicated abortion (10%) [11].

The 2000 DHS showed only 27% of the mothers received ANC from health professional and less than one percent of mothers received ANC from traditional birth attendants. The proportion of institutional deliveries was also low for all the regions of the country with only eight or less percents delivering in health facilities [11].

The 2011 EDHS report show that 34% of women who gave birth in the five years preceding the survey received antenatal care from a trained health professional at least once for their last birth. Antenatal care from a trained health professional has increased by 6 percent since the 2005 EDHS estimate (28%) [12].

Access to proper medical attention and hygienic conditions during delivery can reduce the risk of complications and infections that may lead to death or serious illness for the mother and/or baby [13, 14].

1.2 Statement of the Problem

Utilization of maternal health care services is an outcome of social process in which both the social characteristics of an individual such as social class, and structural characteristics, such as the availability and accessibility of health service play a role [15]. The pattern of association between individual (woman's) socio-demographic characteristics and the utilization of maternal health care services, however, depend on the social context of a given society. Of the structural characteristics, distance from the available maternal health services and the costs of these services were often mentioned as obstacles to utilization [15].

In Ethiopia, only few studies have been undertaken concerning factors that influence the utilization of maternal health services in the country. One of the researches was conducted by Mekonnen Yared and Asnakech Mekonnen [6] to study the utilization of maternal healthcare service in Ethiopia by elucidating the various factors influencing the use of these services in the country using 2000 EDHS dataset.

Most of the studies done in Ethiopia so far used statistical techniques such as regression analysis on a limited set of data to assess the factors which contribute to the underutilization of maternal health care service utilization. Since the analysis made by using these methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world databases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases.

Therefore, the goal of this study is to apply data mining classification techniques to discover maternal health care seeking patterns. Identifying the underlying factors that affect maternal health care seeking pattern in the country is also the purpose of this study. The researcher considered selective techniques and tools which will predict maternal health care seeking behavior using EDHS 2011 dataset.

To assess the problem the following research question is formulated for investigation:

- What are the major factors that affect maternal health care seeking behaviour in Ethiopia?
- Which data mining technique is more appropriate to construct maternal health care seeking pattern model that can be used for prediction?

1.3 Objectives of the Study

1.3.1 General Objectives

- The general objective of the research was to develop a predictive model for maternal health care seeking pattern using data mining techniques among women aged 15-49 in Ethiopia.

1.3.2 Specific Objectives

In order to accomplish the general objective, this study will carry out the following specific objectives:

- To conduct a literature review into the factors affecting utilization of Maternal Health care seeking,
- To identify the major factors affecting maternal health care seeking pattern,
- To prepare good quality dataset for data mining techniques,
- To develop a model that can predict the maternal health care seeking pattern of women of reproductive age,
- To report research findings and to make recommendations,

1.4 Scope and Limitation of the Study

The scope of this research was limited to predicting maternal health care seeking pattern using 2011 EDHS dataset among women of reproductive age group. The study was also restricted to use the socio-economic and demographic factors to develop the model. The unavailability of related literature using data mining in the area was one of the limitations encountered to undertake the study. Furthermore, the study does not incorporate the deployment of the model.

1.5 Significance of the Study

The Government of Ethiopia is committed to achieving the MDG5, to improve maternal health, with a target of reducing the MMR by three-quarters over the period 1990 to 2015. Accordingly, the FMOH has applied multi-pronged approaches to reducing maternal and new-born morbidity and mortality. Improving access to and strengthening facility-based maternal and new-born services is one such approach, and is also a Health Sector Development Plan (HSDP) strategic objective as reported in 2011 EDHS.

Among the core elements of the Ethiopian health policy are democratization and decentralization of the health care system; development of the preventable, promotive and curative components of the health care; assurance of accessibility of health care for all

segments of the population; promotion of private sector and NGOs participation in the health sector.

The FMOH has formulated and implemented a number of policies and strategies that afforded an effective framework for improving health in the country including maternal and neonatal health such as making pregnancy safer, reproductive health strategy, youth reproductive health strategy, health care financing, etc.

Data mining technology provides a user oriented approach to find novel and hidden patterns in the data that can be used for decision making. The discovered patterns (rules) can be used by the health care administrators to improve the quality of service. Predicting maternal health care seeking pattern and extracting the knowledge from 2011 EDHS dataset is very helpful to improve maternal and neonatal health.

The findings of the study can support the national health care policy in revising the existing rules by introducing new rules. It is also possible to use the research result, the predictive model as a framework for improving maternal health care. Therefore, the model can be used by the stakeholders to take the necessary action to improve maternal health status of women of reproductive age in the country.

1.6 Methods

1.6.1 Research Design

In order to achieve the above stated objectives, the researcher has used Hybrid methodology of KDP (Knowledge Discovery Process) to build a predictive model using data mining techniques. Hybrid Process model is selected since it combines best features of CRISP (Cross – Industry Standard Process for Data Mining) and KDD (Knowledge Discovery in Database) methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives.

KDP is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [16].

Therefore, the overall research design was to build a model that can be used to predict the maternal health care seeking pattern on EDHS 2011 dataset. One of the most important aspects of this model is iterative and interactive feature. The feedback loops are necessary because any changes and decisions made in one of the steps can result in changes in subsequent steps.

The six steps of hybrid KDP process are employed in this study are as follows:

1.6.2 Understanding the Business Domain

The first step in the hybrid KDP process is to understand the problem domain to define the problem and determine the research goals and learning about current solution to the problem. It also involves learning domain – specific terminology and preparation of a description of the problem, including its restriction. The United States Agency for International development (USAID) introduced the DHS programme in 1984. Since the establishment of this program has helped more than 85 countries by giving technical support in conducting more than 240 surveys. This program has gained reputation in collecting accurate and representative data throughout the world. The survey is primarily designed to collect data on marriage, fertility, family planning, maternal and child health, HIV/AIDS, malaria, nutrition and gender based violence and traditional malpractices. In relation to this woman aged 15-49 years and men aged 15-59 are the main focus of this survey. The DHS surveys are typically conducted every five years and usually based on a representative sample size at national, regional and urban-rural residence type.

The 2011 EDHS collected information from nationally representative samples of 16,515 women aged 15-49 years and 14,110 men aged 15-59 years. All these individuals were interviewed during the survey using structured questionnaires. Three questionnaires were used for the interview; household's questionnaire, men's questionnaire and women's questionnaire. The different stakeholders in this survey were ICF International through its MEASURE DHS project, national organizations like ministry of health of Ethiopia, Central Statistical Agency and the Ethiopian Health and Nutrition Research institute (EHNRI).

The 2011 Ethiopia Demographic and Health Survey (EDHS) were carried out under the aegis of the Ministry of Health (MOH) and were implemented by the Central Statistical Agency (CSA). The testing of the blood samples for HIV status was handled by the Ethiopia Health and Nutrition Research Institute (EHNRI). ICF International provided technical assistance as well as funding to the project through the MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide. Funding for the EDHS was provided by the government of Ethiopia and various international donor organizations and governments: the United States Agency for International Development (USAID), the HIV/ AIDS Prevention and Control Office (HAPCO), the United Nations Population Fund (UNFPA), the United Nations Children's Fund (UNICEF), the United Kingdom Department for International Development (DFID), and the United States Centers for Disease Control and Prevention (CDC). For this

particular study women's of reproductive age dataset collected on maternal health care utilization from the EDHS 2011 was used. Finally the researcher translated the goals into data mining objectives and initial selection of data to be used was performed in the next step.

1.6.3 Understanding the Data

The source of data for this research was 2011 EDHS dataset available from CSA or www.measuredhs.com after getting consent from the organizations. 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. Among these EDHS has collected high quality data on family health, including immunization coverage among children, prevalence and treatment of diarrhoea and other diseases among children under age five, and maternity care indicators, including antenatal care, assistance at delivery and postnatal care. To understand the data, discussions were made with domain expert from FMOH (EPI and MCH Focal person) and EDHS reports, manuals and brochures were also thoroughly explored.

The sampling in the 2011 EDHS is stratified, clustered and selected in two stages. In the first stage, 624 clusters were selected (187 urban and 437 rural) from the list of enumeration areas from the 2007 census sampling frame. In the second stage a complete household listing was carried out in each selected clusters. This gave 17,817 households with representative sample of 16,515 women between the ages of 15-49 and 14,110 men between the ages of 15-59. Therefore, the researcher has used data collected from women's of reproductive age of number 16,515. The 2011 EDHS collected information on ANC coverage, Assistance during delivery and Postnatal care from responses of women who had a live birth in the five years preceding the survey. For women with two or more live births during the five-year period, the EDHS data refer to the most recent birth only for ANC coverage. Therefore, a sample size of 7,764 women's was used in order to build a quality model that will be used for prediction.

1.6.4 Data Preprocessing

The third step is data preparation which concerns deciding about the data that will be used as input for DM methods in the subsequent steps. After the selection of relevant features with the help of the domain expert and extensive literature survey, the researcher has selected women's individual dataset based on the objective of the study. Then the next step was data preparation used to make ready the selected data for mining purpose. In this phase data were

cleansed such as handling missing values, data integration, and transformation and finally data was converted to the necessary format for mining software as described in detail in Chapter four. The results are data that meet the specific input requirements for DM tools.

1.6.5 Data Mining

The fourth step is data mining that data miner uses various data mining techniques such as classification, clustering and association rule discovery to derive hidden knowledge from processed data. This step creates predictive and/or descriptive models.

Weka 3.6.8 machine learning software was used for analyzing the preprocessed data to build the model. Decision tree and Naïve Bayes algorithm were used among the available algorithms due to their popularity in the recently published papers and due to the objective of the research. For this reason the researcher has used data mining techniques J48 decision tree and Naïve Bayes to build the model for predicting maternal health care seeking pattern using EDHS 2011 dataset.

1.6.6 Evaluation of the Discovered Knowledge

The discovered knowledge is evaluated for understanding the result, checking whether the discovered knowledge is novel and interesting. Interpretation of the results by domain experts and checking the impact of the discovered knowledge is part of this phase. Thus after the development of the model based on the training dataset, the accuracy of the model were tested using test datasets. A confusion matrix and the result of the model were assessed to compare the discovered knowledge.

1.6.7 Use of the Discovered Knowledge

Finally there is a need to plan where and how to use the discovered knowledge. A plan to monitor the implementation of the discovered knowledge will be created and the entire research will be documented.

The result of this study will be presented and disseminated to different organization/bodies such as Addis Ababa University, School of Information Science and Public Health submitting original copy and every effort will be made to disseminate the results of the study through the following ways:

- Presentation for the school of public Health and school of Information Science
- Publishing in different journals
- Presentation on different conferences/workshops for the concerned bodies

- Putting the hardcopy in the libraries of concerned organizations so that interested readers can get access to the research output to be used for decision support, take action or to use it as base for further research in the area

1.7 Ethical Consideration

All the data used for the study were publicly available or made available upon request from the relevant agencies for academic purpose. Furthermore, the data were in aggregate form and no data was collected at the individual level; hence, anonymity or confidentially issues did not arise.

1.8 Organization of the Thesis

The research is presented in six chapters. Chapter one highlights major issue relating to maternal health at a global level and Ethiopia in particular. The objectives and significance of the study are also described. In addition, it covers the methodological issues and the sampling procedure.

Chapter two contains of literature review relating to factors associated with the utilization of maternal health care services. It reviews also data mining concepts and its application in health care endeavor. Chapter three discusses the Mining techniques and methods used for creating the classification models using Decision tree and Naïve Bayes algorithms. The fourth Chapter describes business understanding and data selection. Hence, it gives the statistical summary of the dataset used for the research. In Chapter five deals with the experiment of the research. It comprises modeling, evaluation, presentation and interpretation of the rules discovered from the experiments. Finally, in Chapter six conclusion and recommendations of the study are presented.

CHAPTER TWO

LITERATURE REVIEW

Nowadays, digital information is relatively easy to capture and fairly inexpensive to store. The digital revolution has seen collections of data grow in size, and the complexity of the data therein increase. Advances in technology have resulted in our ability to meaningfully analyze and understand the data gathered lagging far behind our ability to capture and store these data [17]. A question commonly arising as a result of this state of affairs is having gathered such quantities of data, what do actually do with it [18]?

It is often the case that large collections of data, however well structured, conceal implicit patterns of information that cannot be readily detected by conventional analysis techniques [19]. Such information may often be usefully analyzed using a set of techniques referred to as knowledge discovery or data mining. These techniques essentially seek to build a better understanding of data, and in building characterizations of data that can be used as a basis for further analysis [20], extract value from volume [21].

2.1 Data Mining and Knowledge Discovery from Data (KDD)

It is generally accepted that the reason for capturing and storing large amounts of data is due to the belief that there is valuable information implicitly coded within it [18]. An important issue is therefore how is this hidden information (if it exists at all) to be revealed? Traditional methods of knowledge generation rely largely upon manual analysis and interpretation [17]. However, as data collections continue to grow in size and complexity, there is a corresponding growing need for more sophisticated techniques of analysis [17]. One such innovative approach to the knowledge discovery process is known as Data Mining.

Data mining is essentially the computer-assisted process of information analysis [20]. This can be performed using either a top-down or a bottom-up approach. Bottom-up data mining analyses raw data in an attempt to discover hidden trends and groups, whereas the aim of top-down data mining is to test a specific hypothesis [22]. Data mining may be performed using a variety of techniques, including intelligent agents, powerful database queries, and multi-dimensional analysis tools [23].

The data mining approach expedites the initial stages of information analysis, there by quickly providing initial feedback that may be further and more thoroughly investigated if appropriate. The results obtained are not (unless otherwise specified) influenced by

preconceptions of the semantics of the data undergoing analysis. Patterns and trends may therefore be revealed that may otherwise remain undetected, and/or not considered [24].

2.2 Data Mining, Machine Learning and Statistics

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification. Both communities have made great contributions to the understanding and application of neural nets and decision trees [25].

Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions [25].

New techniques include relatively recent algorithms like neural nets and decision trees, and new approaches to older algorithms such as discriminant analysis. By virtue of bringing to bear the increased computer power on the huge volumes of available data, these techniques can approximate almost any functional form or interaction on their own. Traditional statistical techniques rely on the modeler to specify the functional form and interactions [25].

The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. Data mining is a tool for increasing the productivity of people trying to build predictive models [25].

2.3 Knowledge Discovery Process

Knowledge Discovery Process is defined as “The nontrivial extraction of implicit, previously unknown, and potentially useful information from data” [25].

According to researchers such as Fayyad, *et al*, the process of knowledge discovery via data mining can be divided into four basic activities; *selection*, *pre-processing*, *data mining*, and *interpretation*. These stages are discussed in the following section. A graphical representation of the general process is presented in Figure 2.1.

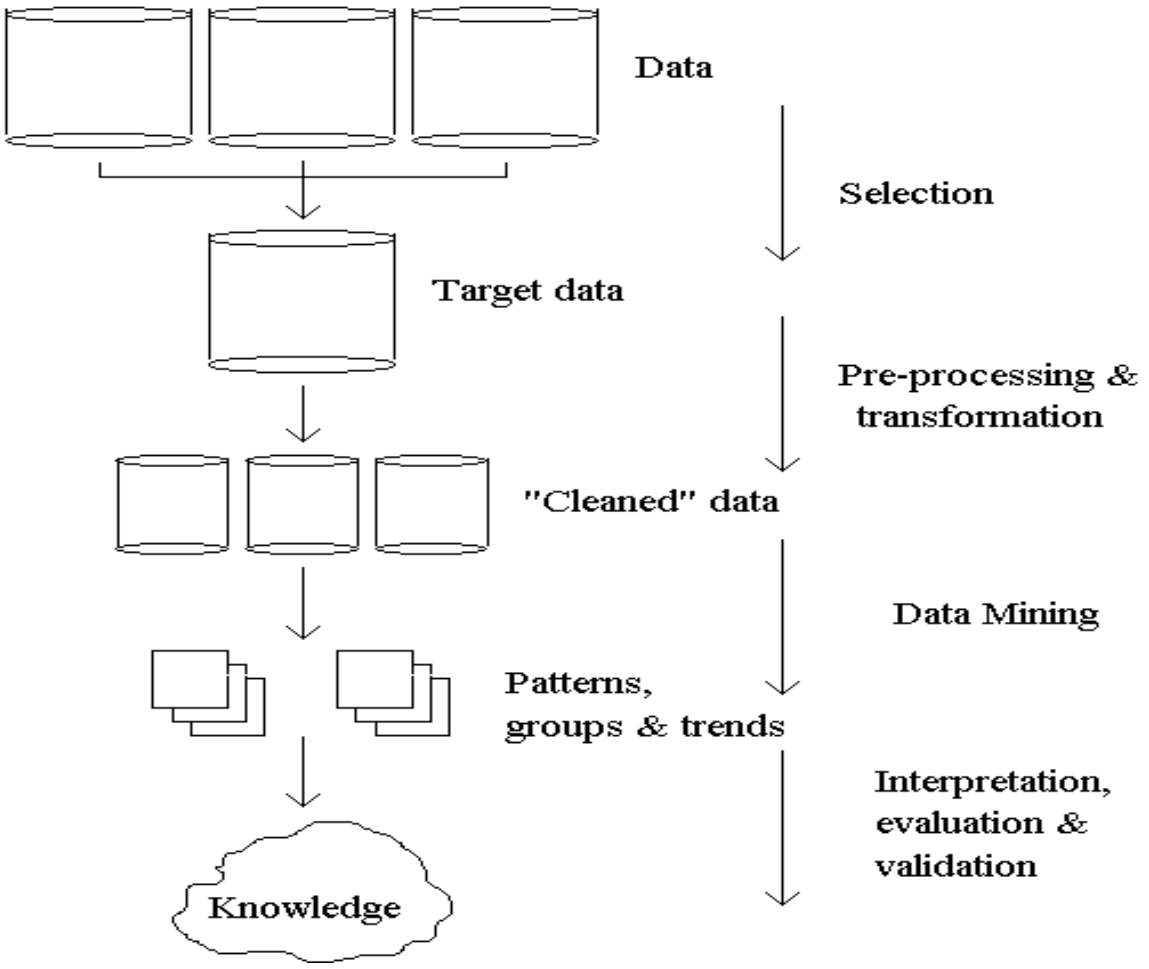


Figure 2.1: Knowledge Discovery Overview

Selection

Selection involves creating the target data set, i.e. the data set to about undergo analysis. As discussed previously, modern datasets may be both large and complex. Large datasets which are not particularly complex may generally be subjected in their entirety to the analysis process (subject to technical constraints). Indeed, the larger the amount of available data, the greater the likelihood that an identifiable trend or pattern may be identified and statistically validated [25].

However, if the dataset is relatively complex, it is often considered impractical to attempt to subject the complete dataset for analysis. It is a common misconception to assume that the complete dataset should be submitted to the data mining software, which in turn will automatically resolve any problems and make sense of any inconsistencies. This is not in fact the case, and is partly due to the probability that the data represents a number of different aspects of the domain which may not be directly related. Subjecting such data to automated analysis may result in the identification of meaningless patterns or trends, which in turn

wastes time and effort. Careful thought should therefore be given as to the purpose of the analysis exercise, and a target dataset created which contains data that reflects this purpose [25].

Pre-processing

Pre-processing involves preparing the dataset for analysis by the data mining software to be used. This may involve resolving undesirable data characteristics such as missing data (non-complete fields), irrelevant fields, non-variant fields, skewed fields, and outlying data points. The pre-processing activities may result in the generation of a number of (potentially overlapping) subsets of the original target dataset.

Data fields are generally viewed as being complete if 70% or more of the records contain values [26]. In cases where the field is considered theoretically complete, but in fact is less than 100% complete, various techniques such as estimation or assigning the category mode are available for producing *synthetic* data. The generation of accurate values to represent missing data is currently one of the main research areas occupying the data mining community [17].

Depending upon the original source of the data, and the storage format employed, pre-processing may also involve converting the data into a format acceptable to the data mining software being used. This initial collection and manipulation of data (during the selection and pre-processing stages) in the data mining process is sometimes referred to as collection and cleaning.

Data mining

This involves subjecting the cleaned data to analysis by the data mining software in an attempt to identify hidden trends or patterns, or to test specific hypotheses. It is recommended that any (apparently) significant results obtained are validated using traditional statistical techniques at this stage.

Evaluation and Interpretation

This involves the analysis and interpretation of the results produced. This may well involve returning to previous stages to carry out additional activities in order to provide further information if necessary.

2.4 Data Mining Tasks

Data mining tasks are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories, Descriptive and Predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions [27].

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined [28]. Descriptive models belong to the realm of unsupervised learning. Such models interrogate the database to identify patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, among others, belong to this family of descriptive models [27].

In predictive modeling tasks, one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models such as classification, regression and Artificial Intelligence -based models. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. Descriptive techniques are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms [29].

2.4.1 Descriptive Modeling Techniques

It is the summary of the data in a convenient and concise way that can be represented in the form of numeric or graphically in a form that can be interpretable by human. It focuses on finding human-interpretable patterns describing the data. The following are the major techniques used by this method:

2.4.1.1 Clustering

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures [27]. According to them cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering. In this context, different clustering methods may generate different clustering's on the same data set. The partitioning is not performed by humans, but

by the clustering algorithm. Hence, clustering is useful in that it can lead to the discovery of previously unknown groups within the data.

Unlike classification in clustering there is no predefined (labeled) class and records are grouped together on the basis of similarity among the instances.

2.4.1.2 Summarization

It is also called data characterization used to summarize the general characteristics or features of a target class of data. There are several methods for effective data summarization and characterization. Simple data summaries based on statistical measures and plots are used to present the data in terms of numbers or visually by graphs.

2.4.1.3 Association Rule Discovery

Association rules are one of the major techniques of data mining in unsupervised learning system. While classification and clustering are global pattern discovery of DM task, association rule discovery is the most common form of local-pattern discovery [30]. Of the data mining tasks, association rule mining method is applied in either supervised or unsupervised manner [30]. Association rule mining searches for interesting relationships among items in a given dataset and patterns are represented in the form of association rules [31]. Association rules such as the occurrence of some items in a transaction will imply occurrence of other items in the same transactions [31].

Interestingness of patterns needs to be measured to make sense that patterns are easily understood by humans, valid, potentially useful, and novel [31]. In the case of classification rules, we are generally interested in the quality of a rule set as a whole. It is all the rules working in combination that determine the effectiveness of a classifier, not any individual rule but, in the case of association rule mining the emphasis is on the quality of each individual rule [31].

2.4.2 Predictive Modeling

It involves using some variables or fields in the database to predict unknown or future values of other variables of interest. The following are the major techniques used for predictive modeling:

2.4.2.1 Classification

Classification is one of the predictive data mining tasks. It is a technique used to predict group membership for data instances by assigning previously unseen records a class as

accurately as possible. It is said to be the process of finding a model or function that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown [31].

The derived model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision trees, mathematical formulae, semantic network etc [27]. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attributes set and the class level of the input data.

After having an accepted accuracy level, one can use the model for classification of new data set.

2.4.2.1.1 Decision Tree Induction

When decision tree induction is used for attribute subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples. Thus, the algorithm choose the best attribute to partition the data into individual classes includes ID3, C4.5, and CART [31].

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

There should be certain requirements before decision tree algorithms become applied [31]. At first decision tree algorithms represent supervised learning; they require pre-defined target variables and training dataset which provides the algorithm with the values of the target variable.

Then this training dataset should be rich and varied, providing the algorithm with a healthy cross-section of the types of records for which classification may be needed in the future.

Decision trees learn by example (training sample), and if examples are systematically lacking for a definable subset of records, classification and prediction for this subset will be problematic or impossible.

Finally the target attribute classes must be discrete i.e. one cannot apply decision tree analysis to a continuous target variable. The target variable needs to take on values that are clearly demarcated as either belonging or not belonging to a particular class.

One of the most attractive aspects of decision trees lies in their interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf [31]. Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule [31].

The challenge with decision tree is over fitting. As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex [31]. This potentially leads to the concept of over fitting which consequently brings the notion of pruning; this implies removing of branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree [31].

2.4.2.1.2 Naïve Bayes Classifier

Bayesian classifier is statistical classifier and a practical learning algorithm that can predict class membership probabilities. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes and classification is based on a probabilistic model specification; i.e. it can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class [31].

When using the Naïve Bayes method to classify a series of unseen instances the most efficient way to start is by calculating all the prior probabilities and also all the conditional probabilities involving one attribute though not all of them may be required for classifying any particular instance. Naïve Bayesian classifier assumes class conditional independence i.e. it treats each variable independently and measure the effect that different values of the variables [31].

The advantage of Naïve Bayes classifier is it reaches the minimum error when the dataset is large and the methods for estimating particular probabilities are consistent [31]. However, the challenge in Naïve Bayes classifier is that the model used in the classification might not be the best estimator of the probability distribution though it has multi-effects in different area where it has relatively good performance [31].

In Bayesian network (Belief network) which is the graphical model of causal relationships that represent dependency among the variables and it gives a specification of joint probability distribution so that it is used to solve the variables interdependences [31].

2.5 Data Mining Methodologies

There are different types of processes that are used by data mining to extract useful information from a dataset. These are the major methods used:

2.5.2 Knowledge Discovery in Database (KDD)

The first KDD process was proposed by Fayyad in 1996 [16]. This process consists of several steps that can be executed iteratively. KDD has been more formally defined as it is non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is the process of knowledge discovery while data mining is a technique applied for knowledge discovery considered as just a step in the entire process [16]. As shown in Figure 2.2, the KDD process consists of five steps: Data Selection, Data Pre-processing, Data Transformation, Data Mining and Interpretation/Evaluation.

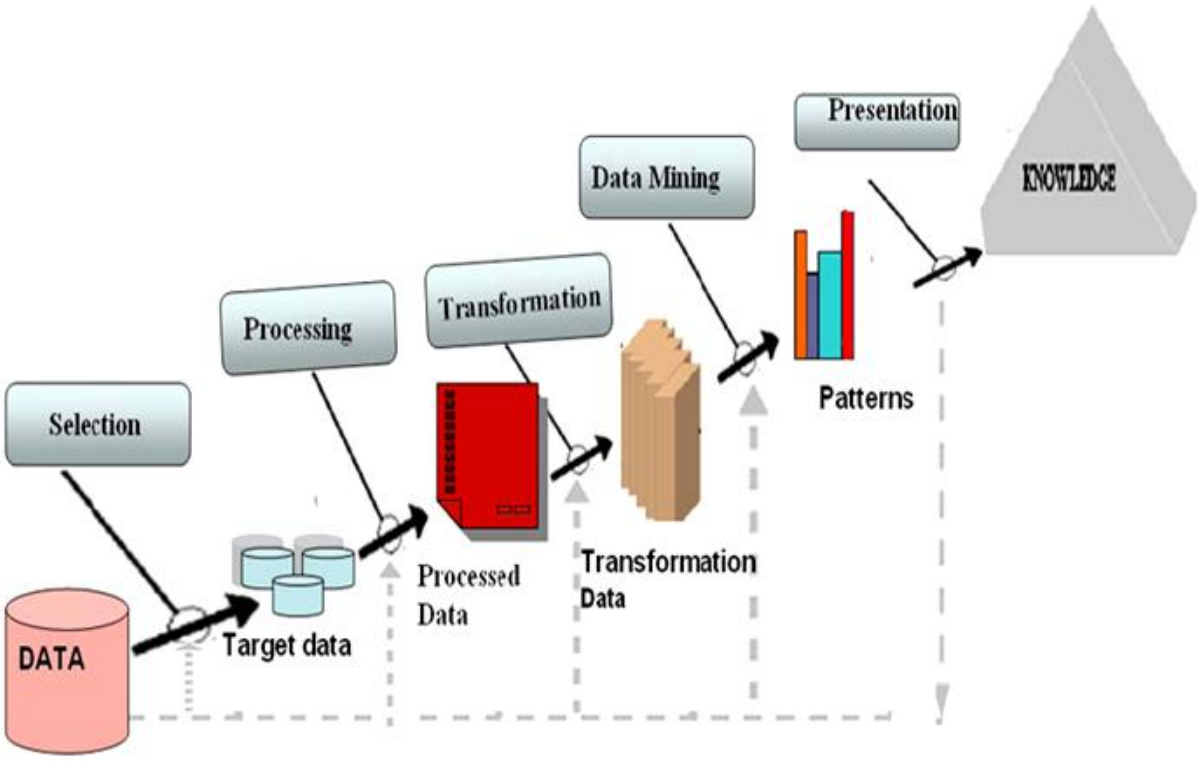


Figure 2.2: Knowledge Discovery Process

2.5.3 CRISP

CRISP–DM was developed in 1996 by analysts for fitting data mining into the general problem solving strategy of a business or research unit [32]. CRISP–DM is one of the most widely used methodologies in extraction of knowledge which has a life cycle consisting of six phases which is an iterative and adaptive process [32], as depicted in Figure 2.3.

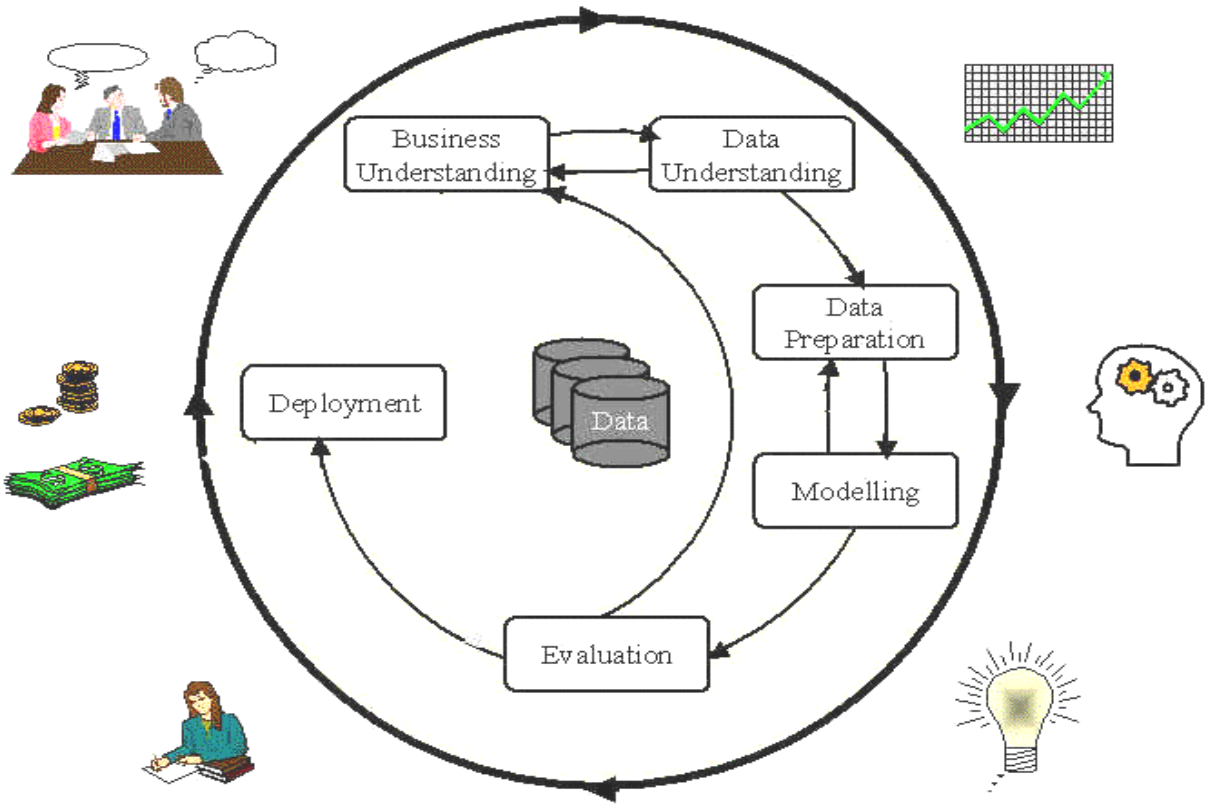


Figure 2.3:The CRISP-DM knowledge discovery Process Model

2.5.4 Hybrid Model

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

Hybrid process is characterized by providing more general, research oriented description of the steps. The hybrid model also encourages the application of knowledge discovered for a particular domain in other domains and it has a six step process as depicted in Figure 2.4.

The six step in hybrid KDP model as shown in the fig below:

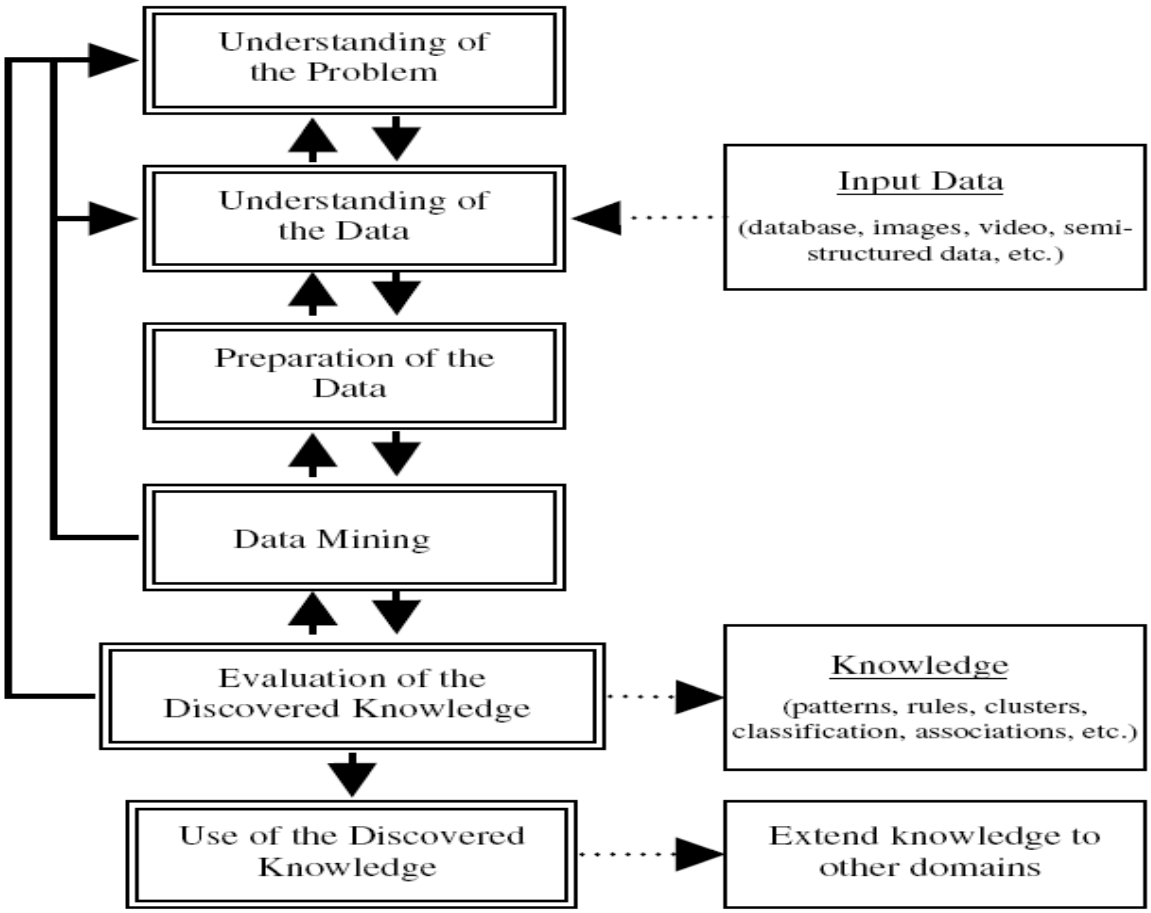


Figure 2.4: Hybrid Process Models [34]

One of the most important aspects of this model is iterative and interactive feature. The feedback loops are necessary because any changes and decisions made in one of the steps can result in changes in subsequent steps.

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

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

The third step is data preparation which concerns deciding about the data that is used as input for DM methods in the subsequent steps. The cleaned data may be further processed by feature selection and extraction algorithms.

The fourth step in data mining that data miner uses various data mining techniques such as classification, clustering and association rule discovery to derive hidden knowledge from processed data. This step creates Predictive and/or Descriptive models. The discovered knowledge is evaluated for understanding the result, checking whether the discovered knowledge is novel and interesting. Interpretation of the results by domain experts and checking the impact of the discovered knowledge is part of this phase.

2.5.5 SEMMA

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA stands for (Sample, Explore, Modify, Model, and Assess) refers to a methodology that clarifies this process [34]. The SAS Institute considers a cycle with five stages for the process:

Sample – this stage consists on 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.

Explore - this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify - this stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model - this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals [35].

2.6 Application of Data Mining in Health Care

There is vast potential for data mining applications in healthcare. Generally, these can be grouped as the Evaluation of treatment effectiveness; Management of Healthcare; Customer relationship management; and Detection of fraud and abuse etc. [36].

2.7 Related Works

Providing special care for pregnant women in the public health services was not started until 1930s, which was initially introduced in the United Kingdom and Northern Ireland states. It was decided that every pregnant woman should get a regular check-up as an integral part of maternity care. This is one of the important components of maternal health which is now called antenatal care (ANC). Antenatal service is important as it offers pregnant women an opportunity to get different services which alerts the woman to the risks associated with pregnancy and for discussing her options for safe delivery [37, 38].

The main contributory factor for the development of this special care for pregnant women was the realization of increased occurrences of deaths due to eclampsia, while deaths due to the other common causes of maternal death like sepsis, haemorrhage and obstructed labour started to decrease during the 20th century. In order to decrease deaths due to eclampsia, it was necessary to detect the problem early during pregnancy by measuring blood pressure and by identifying those women at higher risk for convulsion and starting early intervention to decrease blood pressure when possible [37].

Ideally ANC should consist of health education for pregnant women, early screening to identify those at high risk of developing complications and diagnosing problems if there are any. Whenever possible it is also important to intervene in order to prevent the development of complications. The diagnosis and treatment should not be limited to those conditions that developed during pregnancy but also for any pre-existing medical condition. These actions should lead to a decrease in the risk or severity of morbidity and mortality [39].

Evidences showed the health benefits that can be obtained from some specific services of ANC like tetanus immunization, prevention and treatment of malaria, management of anaemia during pregnancy, and treatment of sexually transmitted infections. Use of these services helps to improve foetal outcomes and maternal health. Recently ANC has become important as an entry point for HIV prevention and care including prevention of mother to child transmission of HIV [40]. Since women who have received ANC are more likely to seek assistance during delivery from a health professional, a focused ANC model in addition to its direct contribution to better health can also contribute to safe delivery [41].

Globally 30% of women aged 15-40 do not have ANC. Forty six percent of those who did not have ANC are in south Asia while 34% are in sub-Saharan Africa. This low use of services leads to death and disability due to untreated hypertensive disorders or due to under-nutrition like iron deficiency anaemia [42]. There has been a significant increase in antenatal service

use between the years 1990-2000; the increase has been more than 20% in all the regions of the world except the sub-Saharan regions where only 4% increase was noted [43].

Regarding delivery care one third of births take place at home without receiving assistance from a skilled birth attendant [44]. A skilled attendant of delivery is defined according to the WHO as an accredited health professional – such as a midwife, doctor or nurse – who has been educated and trained to proficiency in the skills needed to manage normal (uncomplicated) pregnancies, childbirth and the immediate postnatal period, and in the identification, management and referral of complications in women and new-born [45].

In countries like Sweden and UK significant numbers of deliveries take place at home by trained midwives with physicians attending only women with complications or who are at high risk. On the other hand in developing countries most women deliver at home and if the woman has complications it may not be detected and she may or may not be taken to a health facility.

These home deliveries may be attended by various health workers, untrained family member or others may deliver completely alone [46].

There has been a significant change in the number of births attended by skilled professionals. In developing countries the proportion of births attended by skilled birth attendants increased from 42% to 53% over the decade from 1990 to 2000. In this period in Asia the births attended by skilled professionals reached 35% while on the other hand the change in the sub-Saharan regions was only 5% [42].

Availability, quality and affordability of maternal health care services for sure influence use of the services by women. But good supply doesn't create demand by itself. Even under same circumstances some women use the services more than the others. This shows that there are factors other than the health care service characteristics that influence the use of maternal health care services [47].

Several studies have shown that socio-demographic factors affect utilization of maternal health care services [47, 50] Some studies showed women's education increases the use of maternal health services [47, 48]. Educated women are more likely than are uneducated women to use ANC, to use it early and frequently, and to use trained providers and medical institutions, similarly education is positively associated with safe delivery. Female education was also seen to be a strong predictor of maternal mortality independent of income per head [47]. In one study from Nepal women with more than primary level education were more

likely to use ANC than those with no education [47]. In a study conducted in Turkey based on the 1993 Turkey's DHS to assess the socio-demographic determinants of maternal health care utilization women with six or more years of schooling were more likely to use ANC than women with no schooling [48].

Place of residence is the other factor that was documented to significantly influence the use of maternal health care services. Rural women are generally less likely to give birth in health facility than their urban counterparts [51, 52]. A systematic review which assessed the inequalities in maternal health service utilization using 30 papers from 23 countries including Ethiopia showed that pattern of use of the maternal health services was different among countries and even within countries. Urban and wealthy women were more likely to deliver with assistance of health professional than rural and poor women. The study also showed that wealthier women were likely to seek early ANC than poor ones [53]. A study done based on the 2000 Ethiopian DHS demonstrated that 27% of mothers who gave birth in the five years before the survey received ANC from health professional and further analysis showed that urban women showed higher use of ANC than the rural counterparts, 83% of women in Addis to 22% women in the rural regions [50].

In another study conducted in the Northwest part of Ethiopia (Gondar) to asses safe delivery service utilization among women of childbearing age, 46% of the women attended ANC at least once in their pregnancy, the percentage of women living in urban area and receiving ANC was about three times higher than those mothers living in rural parts of the region. Only 14% of the mothers gave birth in health facilities out of this 2% of women living in the rural regions gave birth in health facilities [51].

Economic stability of households is also one of the well-recognized factors that can affect the utilization behaviour of a woman. The poorest women in the poorest regions of the world have the lowest service coverage. A study in over 50 countries showed that on average more than 80% of births were attended for the richest women compared with only 34% of the poorest women [43].

In a study from Nepal, household economic status in particular was found to be an important factor associated with utilization of maternal health care services. This can be explained by the ability to pay for services by economically well off groups but the fact that there was a significant relationship after controlling for other factors like place of residence suggests that the richest groups differ from their poor counterparts by more than just dispensable income [47]. Women's economic opportunity in providing for the family measured by their

involvement in gainful or paid employment, type of occupation and status of work also affects their health and health seeking behaviour. This might empower women and they will have increased control over income and on decision making concerning their health. As a result they will have increased health seeking behaviour leading to improved maternal health [43].

On the other hand employment may also pose physical exhaustion and in some cases employed women may not have the time to go to health services this may have a negative effect on use of the services [43, 47]. Religion is the other variable which was seen to have some significant relation with service utilization. In a study from the 2000 EDHS it was noticed that those individuals following Orthodox/Catholic, Muslim and Protestant tend to use ANC more than those following traditional belief [50]. Women having children with birth order of five or more and who is grand-multipara have a lesser chance of delivering in health institutions than those with lesser number of children [54].

Chen et al. [55] investigate the possible effects of multiple drug exposures at different stages of pregnancy on preterm birth, using SmartRule, a data mining technique for generating associative rules. In this work, two subsets of Danish National Birth Cohort (DNBC) dataset are used. The first subset contains 4454 records including 1000 women who were depressed and/or exposed to various active drugs. This set is used for finding the side effects of anti-depression drugs. The second subset contains 6231 records, including 414 preterm cases. This set is used for finding side effects of multiple types of drugs. The authors develop a tree hierarchical model for organizing the generated rules, in order to ease the recognition of interesting rules by human experts. Using this system, the authors claim that they are able to find novel and interesting rules.

In a study conducted by Biset Desaleng [56], with a title "predicting low birth weight using data mining techniques on Ethiopia Demographic and Health Survey Data Sets" has employed Knowledge Discovery Process using CRISP and KDD methodology to build a predictive model that can be used to predict the low birth weight of child during delivery. The researcher also used in his study the mining tool WEKA to build a model using J48 decision tree classifier and PART rule induction algorithms were selected to predict children as Low and Normal classes using independent variables.

In general, the results from this study were encouraging; The extracted rules in both the algorithms are very effective for the prediction of low birth weight. It is possible to observe, from both algorithms that the 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.

A variety of factors have been identified as the main causes of utilization of maternal health care services including demographic and socio-economic status, lack of physical accessibility, cultural beliefs and perceptions, low literacy level of the mothers and large family size are commonly found.

The model of utilization of maternal health service that was used in the analysis is based up on the conceptual framework of health-seeking behaviour developed by Anderson and Newman [57]. This behavioural model proposed that the use of health care services is a function of three sets of individual characteristics: (i) predisposing characteristics, e.g. age, household size, education, number of previous pregnancies, health-related attitude; (ii) enabling characteristics, i.e. income, characteristics of health care system and accesses, and availability of health facilities; and (iii) need characteristics, i.e. characteristics of illness, perceived health status, and expected benefit from treatments.

To the best of the knowledge of the present researcher, there was no research done using data mining techniques to predict maternal health care seeking pattern in Ethiopia.

CHAPTER THREE

METHODOLOGY

As previously explained in Chapter two, the methods and techniques of Data mining and its application have been thoroughly discussed. From these the researcher has selected the six-step Hybrid Knowledge Discovery Process model in order to achieve the objective of the research on building predictive models as discussed in Chapter one section 1.6. Since the purpose of the research is to build a predictive model based on the secondary data source taken from EDHS 2011 particularly the researcher has used women's of reproductive age to achieve the objective of the research. Therefore, both Descriptive and Predictive methods have been implemented for building the model. The DM techniques used in this research were J48 decision tree algorithm and Naïve Bayes classification algorithms.

3.1 Problem Domain Understanding

In order to understand the problem domain the researcher has done a thorough investigation on issues that were researched so as to see a gap where the potentials of data mining could be used to fill. The researcher has reviewed studies conducted to identify most significant factors that affect maternal health care seeking behaviour and to predict the most predominant factors for maternal health care seeking pattern with the use of Data Mining techniques.

In addition to what is learned from reviewing the literatures, experts in the profession were identified and contacted in order to further increase the knowledge about business domain. Moreover, efforts exerted to understand the problem domain has given an opportunity to the researcher to learn many maternal health care related terminologies. The knowledge obtained through the above mentioned ways were used in stating the business objectives of this study. Specifically, predicting maternal healthcare seeking pattern also improve the health-seeking behaviour of the client, orient the client to birth preparedness issues, and provide basic preventive and therapeutic care is identified as the business objective of this study, which is translated to the data mining objective stated in the first Chapter of section 1.3.

3.2 Data Understanding and Data Preparation

After understanding the business domain the researcher has made a great deal of effort to understand and pre-process the EDHS 2011 data for applying the selected DM techniques. Tasks such as attribute and dataset selection, handling missing values and data transformation were performed on the selected dataset in order to use it for the Data mining task at hand. It is also concerned on converting from one format to another format to make the dataset for

Weka software understandable format which is used as input for data mining process in subsequent steps. To develop first insight into the data, relevance analysis like descriptive data visualization was also done using statistical tool (SPSS 20) and Excel 2007 were also used.

In this particular research, women's of reproductive age 15-49 who had a live birth in the five years preceding the survey and healthcare received (i.e. Antenatal, Delivery and Postnatal) by skilled provider that is, from a Doctor, Nurse, or Midwife, for their last birth were considered from the total of 16,515 women's individual dataset.

In order to handle different data related problems from the secondary data used from EDHS 2011 dataset like missing values, noises, and irrelevant features were solved using tools like Microsoft Excel 2007, SPSS 20, and Weka 3.6.8. Exporting data from existing format (SPSS 20) to Weka understandable format like arff and csv were also employed.

3.3 Modeling Techniques

These days, data mining is more popular in different areas including healthcare by mining knowledge from the massive dataset. Commonly used techniques for data classification and prediction in data mining are decision tree induction, Bayesian classification, rule-based induction, the neural network support vector machines, and *k*-nearest neighbor classifiers [57]. Hence, for this particular research, classification algorithms such as Decision Tree Induction (J48) and Naïve Bayes classifier are used.

3.3.1 J48 Decision Tree Algorithm

Decision tree algorithms have predictive performance ability and capability to discover patterns in huge datasets and understand ability of the generated rules by human. For example, the acquired knowledge in tree form using decision tree takes less mental strain to understand the path from the root to leaf and one can generate rule from the tree in order to predict the class for unknown records. In addition rule assimilation easily by end users, classification steps of decision tree induction is simple and fast, and also tree construction does not require any domain knowledge [58].

Therefore, Weka 3.6.8 software based decision tree algorithm (J48) was used which is a greedy algorithm i.e. it constructs trees in a top-down recursive or divide-and-conquer manner and orders the class rule sets so as to minimize the number of false-positive error [59]. It uses the concept of information gain or entropy reduction to select the attribute with the highest information gain.

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

$$H(X) = -\sum Pj\log2(pj) \qquad 3.1$$

For an event with probability *p*, the average amount of information in bits required to transmit the result is $-\log_2 p$. For variables with several outcomes, we simply use a weighted sum of the $\log_2 pj$'s, with weights equal to the outcome probabilities. Therefore, the mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows [60].

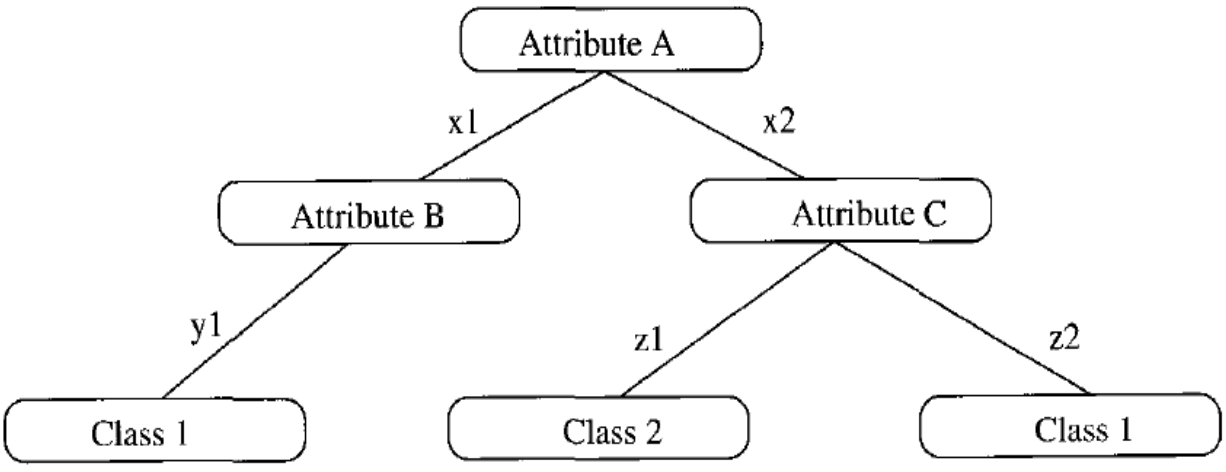
$$\text{Info}_A(X) = \sum_{j=1}^{v} \frac{|X_j|}{|X|} \times \text{Info}(X_j) \qquad 3.2$$

Information gained by branching on attribute A is

$$\text{Gain}(A) = \text{Info}(X) - \text{Info}_A(X) \qquad 3.3$$

At each decision node, C4.5 uses the attribute with the maximum gain ratio as the splitting attribute and recursively visits each decision node, selecting the optimal split, until no further splits are possible. J48 also used the same concept to construct the decision tree and it supports both numeric and nominal predictors and nominal class attribute. It has the capability to handle missing values in datasets [58]. Once the tree is constructed, it is possible to generate the rule in order to apply it for new instances which are independent of the training samples.

Figure 3.1:Simple Decision Tree Constructed for Two Class Classification



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

- Rule 1: If (A = X1 and B = Y1), then Classification = Class 1
- Rule 2: If (A = X2 and C = Z1), then Classification = Class 2
- Rule 3: If (A = X2and C = Z2), then Classification = Class 1

Conditions for stopping partitioning includes all samples for a given node belongs to the same class and there are no remaining attributes for further partitioning.

However, when decision trees are built, many of the branches may reflect noise or outliers in the training data. Considering the goal of research and improving classification accuracy of unseen data, tree pruning was attempted in order to identify and remove unnecessary branches that lead to over-fitting of the model.

The second algorithm is Naïve Bayes classifier which has found to be comparable in performance with other algorithms in data mining like decision tree and neural network classifiers [59]. Naïve Bayesian classifier has also exhibited high accuracy and speed when applied to large databases and it also assumes that the effect of an attribute value on a given class is independent of the values of the other attributes [60].

3.3.2 Naïve Bayes Classifier

Naïve Bayes classifier is one of the algorithms that produces the lowest classification error rate on a validation data and produces high accuracy performance as compared with others algorithms in classification and prediction in data mining [61].

In general, Naïve Bayes follows the following steps for classification purpose: first, Collect data and estimate parameters such as mean and covariance for each class. Second, choose a set of features that a classifier uses to compute a posteriori probability. Third, choose a model to derive a decision rule with these parameters. Fourth, train the classifier to test dataset and classify each sample. Finally evaluate the decision rule in order to improve the choice of features and the overall design of the classifier [60].

3.3.2.1 Bayes Basics Theorem

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

P(H) (prior probability) is the initial probability or the prior probability of each class based on the training tuples and P(X|H) (likelihood) is the probability of observing the sample X,

given that the hypothesis holds or conditional probabilities of attributes value for each class, P(X):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = P(X|H) \times P(H)/P(X) \quad 3.4$$

probability that sample data is observed and hence from the given training data X, posteriori probability of a hypothesis H, P(H|X), follows the Bayes' theorem [59].

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

3.3.2.2 Naïve Bayes Classifier

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

This can be derived from Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad 3.5$$

Since P(X) is constant for all classes, only needs to be maximized

$$P(C_i|X) = P(X|C_i)P(C_i) \quad 3.6$$

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

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

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

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

3.4 Data Mining Tool

In this research Weka 3.6.8 software which is developed at the University of Waikato in New Zealand was used. This software is available for free at www.cs.waikato.ac.nz/ml/Weka website [58]. Weka tool is open-source data mining software in Java with a number of collections of algorithms for data mining tasks, including data pre-processing, association

mining, classification, clustering, and visualization [64]. The tool has graphical user interface which consists of buttons and menu commands and panels on its interface, where each panel is used to perform different tasks. The initiation point in the tool after Weka has been run is graphical user interface with supporting tools. After selecting the application from the menu bar to select the data mining tasks (explorer, experiment, knowledge Flow and simple CLI) by clicking on the option on the menu bar to perform the intended task. In order to start data mining tasks in the Weka, it is necessary to open a dataset from where it is saved and import to Weka software. Once the explorer window is chosen and the application data file is imported, different techniques in the menu bar become active and one can perform different tasks according to proposed data mining objectives.

3.5 Performance Measurement

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

3.6 Confusion Matrix

Confusion matrix is a useful tool for analyzing how well classifier recognized the classes. It is body of table with m by m (row and column) matrix the row corresponds to correct classification and the column corresponds to the predicted classifications. An entry, CMi,j in the first m rows and m columns indicate the number of tuples of class that were labeled by the classifier as class j [59]. For a classifier to have good accuracy, ideally, most of the instances would be represented along the diagonal of the confusion matrix with the rest of the entries being closed to zero [58].

In confusion matrix, there are classifier evaluation metrics like Accuracy, Error rate, Sensitivity, Specificity, Precision, Recall, and F-measure. Table 3.1 shows two class classification result simple confusion matrix which contains both predicted and actual classes.

Table 3.1: Confusion Matrix with Two Classes Classification Result

	Predicted Class		
Actual Class		Class=Yes	Class=No
	Class=Yes	A (TP)	B (FN)
	Class=No	C (FP)	D (TN)

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

Here are some of performance evaluation computational techniques on confusion matrix that are used in this study. Accuracy is the first one which is widely used to check the performance of the model. It is the percentage of test set tuples that are correctly classified [65].

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

The performance of the model enables it to classify the positive cases correctly is sensitivity. It is defined as the probability of having a positive test result among those with a positive diagnosis for the disease or true Positive recognition rate [65].

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

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

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

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

Precision = TP/ TP+FP----------For True Class 3.11

Precision = TN/TN+FN----------For False Class 3.12

Finally, the F measure is the inverse relationship between precision & recall (F_1 or F-score**)**: harmonic mean of precision and recall. It is the point to conclude that the precision and recall of the model are significantly balanced [66].

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

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

3.7 Receiver Operating Characteristic (ROC) Curve

A large number of intelligent medical systems (including medical expert systems, neural networks, classifiers, knowledge discovery and data mining systems) showed great progress and they are being developed, practically to aid clinician and to improve patient care in areas such as diagnosis, prognosis, decision support and screening. To test which classifier is

highly significant for a given subject is determined by ROC analysis and it is becoming widely used tool in medical tests evaluation [66].

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified [69]. The following Figure 3.2 shows the performance of classifier 2; that it has the maximum area under curve [68]

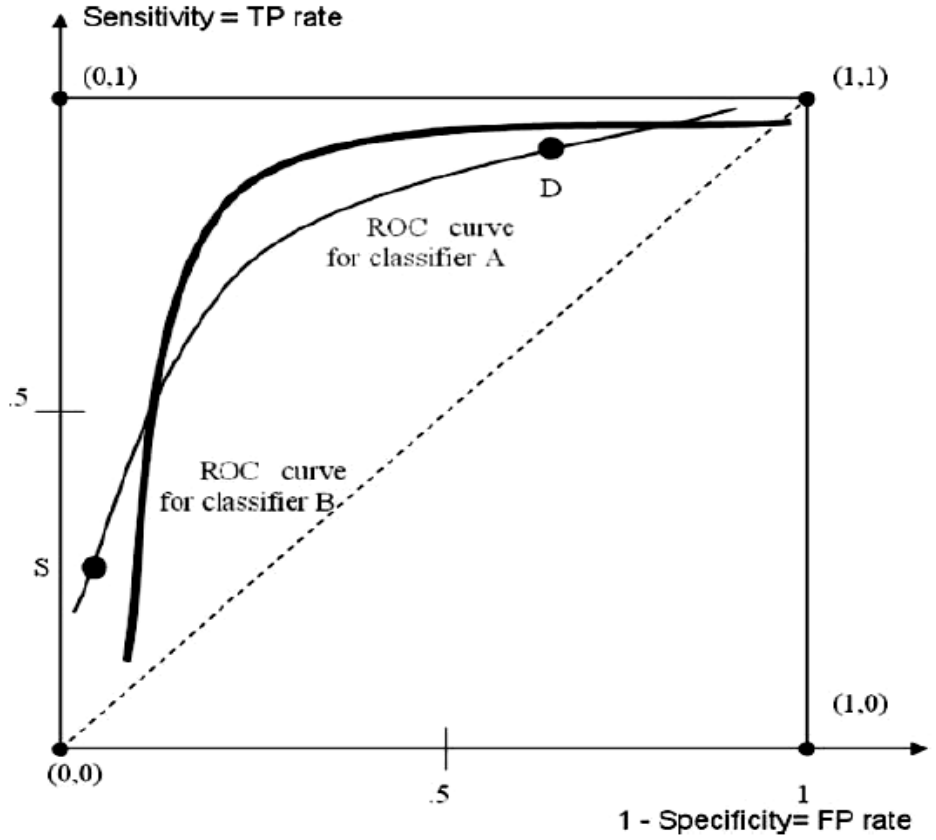


Figure 3.2: Examples for ROC curve

ROC curve is useful visual tool for comparing classification models. It shows the trade-off between the true positive rate (proportion of positive tuples that are correctly identified) and the false-positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model [58, 68]. It is performed by drawing curve in two dimensional spaces by representing vertical axis for true-positive rate and the horizontal axis for false-positive rate [58]. In ROC curve, plotting starts at the bottom left-hand corner where the true positive rate and false-positive rate are zero. To plot a ROC curve for a given classification model, one needs to rank the test tuples in decreasing order.

To assess the accuracy of a model, one can measure the area under the curve which is a portion of the area of the unit square and its value is ranged from 0-1. It is assumed that increasing numbers on the scale represents that the subject belongs to one category while

decreasing numbers on the scale represent the increasing belief that the subject belongs to the other category. Thus, from the ROC curve, the closer the ROC curve of a model is to the diagonal line, the less accurate the model is closer to the area of 0.5.

Table 3.2: **Performance Measures of ROC Area**

No.	ROC Area	Performance
1	0.9-1.0	Excellent
2	0.8-0.9	Good
3	0.7-0.8	Fair
4	0.6-0.7	Poor
5	0.5-0.6	Fail

The model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger evidence for a positive actual state (1.00) [58, 67, and 68]. For example, in Figure 3.2, classifier 1 performs better than 2.

By using ROC analysis one can identify predictors in order to find the one with optimal characteristics and their associated cut-points.

Therefore, for this particular study Accuracy, Sensitivity, Specificity and ROC area are taken in to account when the classifier performance is evaluated to select the best model.

CHAPTER FOUR

DATA PREPARATION AND PRE-PROCESSING

A hybrid data mining method is adapted which guide the researcher to apply data selection and pre-processing in order to extract useful information from the dataset using DM technique. As this is the very important step of the knowledge discovery process, the researcher has made a great deal of effort to understand the business domain which enables to select the major variables in achieving the objectives of the study. After selecting the major variables that affect maternal health care service seeking behaviour with the assistance of the domain expert and by reviewing related literature, the researcher has employed data pre-processing task. Data pre-processing is an essential step in the process of preparing dataset that is appropriate for mining. The purpose of data pre-processing is to clean the noisy data, extract and merge the data from different sources, and then transform and convert the data into a proper format [69]. It is an important step in data mining, because quality decisions must be based on quality data.

This section briefly deals about the data source and selection, data cleaning and data transformation of the data employed in this study.

4.1 Data Source and Selection

The data source for this study is obtained from the organization's internet website address (i.e. www.measuredhs.com) after getting consent from the organization. The 2011 EDHS was conducted by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health. It is the third Demographic and Health Survey (DHS) conducted in Ethiopia, under the worldwide MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide. The survey interviewed a nationally representative population in about 18,500 households. Out of these households, a nationally representative sample of 16,515 women of age 15–49 and 14,110 men of age 15–59 were interviewed. This represents a response rate of 95% for women. The sample design for the 2011 EDHS provides estimates at the national (total, urban, and rural) and regional levels. The data were collected on key indicators relating to family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women’s empowerment, and knowledge of HIV/AIDS [70].

The primary objectives of the 2011 EDHS were to provide up-to-date information for planning, policy formulation, monitoring, and evaluation of population and health programmes in the country.

Since the goal of this study was to classify women according to their likelihood of seeking maternal health care service, the dataset used from women's of reproductive age. The Woman's data 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 collected information on ANC, assistance during delivery and postnatal care coverage from responses of women who had a live birth in the five years preceding the survey. For women with two or more live births during the five-year period, the EDHS data refer to the most recent birth only for ANC. Therefore, a total of **7,764** respondents in the five years preceding the survey for their last birth reported for 2011 EDHS were used for this particular research.

4.2 Variable Selection on Maternal Healthcare Service Utilization

Deciding on the data that will be used for the analysis is based on several criteria, including its relevance to the data mining goals, as well as quality and technical constraints such as limits on data volume or data types [71]. Therefore, in this research the attributes are selected with the help of domain expert and extensive literature review. Because taking all the variables in the database we have, fed them to the data mining tool and find those which are the best predictors may not work very well. One reason is that the time it takes to build a

model increases as the number of variables increases. Another reason is that blindly including extraneous columns can lead to incorrect models [68]. Thus, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts in order to simplify the task of modeling. Accordingly, the following main variables are selected from the women's individual record dataset:

Therefore, Mother's age at birth, Place of Residence, Region, Household Wealth Index, Religion, Highest educational level, Husbands education level, Total children ever born, Frequency of listening to radio, Frequency of reading newspaper, Frequency of watching television, Marital status, Husbands/partners occupation, Mother's occupation, Antenatal care, Delivery care and Postnatal care are the major variables that affect mothers from using the healthcare service as indicated in literature survey. These variables provide the socio-economic and demographic information for each respondent.

4.3 Description of the Selected Attributes

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

Table 4.1: Description of the selected attributes from EDHS 2011 dataset

No.	Attributes	Description	Values	Data type	Count Missing
1.	Mother's age	Current age of mother	Mother age in five year interval group (15-19,20-24,25-29,30-34,35-39,40-44,45-49)	Categorical	0(0.0%)
2.	Residence	Type of place of residence	Urban or Rural	Categorical	0(0.0%)
3.	Region	The 11 administrative region of the country where they live	Tigray, Afar, Amhara, Oromiya, Somali, Benishangul Gumz, Southern National Nationality People (SNNP), Gambela, Harari, Addis Abeba, Dire Dawa	Categorical	0(0.0%)
4.	Women's education level	The highest level of education attained by women	No education, Primary, Secondary, Tertiary (Higher).	Categorical	0(0.0%)
5.	Household Wealth Index	The living standard of the household	Poorest, Poorer, Middle, Richer, Richest	Categorical	0(0.0%)
6.	Religion	Religion of the mother	Orthodox, Muslim,	Categorical	4(0.0%)

			Protestant, Catholic, Traditional, Others		
7.	Marital status	Marital status of the mother during birth	Never married, Married, Living together, Widowed, Divorced, Not living together	Categorical	0(0.0%)
8.	Husbands education level	Education status of a partner	No education, Primary, Secondary, Tertiary (Higher).	Categorical	161(2.1%)
9.	Total children ever born	Number of children of a women including this one	1 child, 2 or 3 child, 4 or 5 child, more than 6 child	Numeric	0(0.0%)
10.	Frequency reading newspaper	How often do a women listen to radio	Not at all, Less than once a week, At least once a week	Categorical	5(0.00%)
11.	Frequency of listening radio	How often do a women reads newspaper	Not at all, Less than once a week, At least once a week	Categorical	5(0.00%)
12.	Frequency of watching television	How often do a women watches television	Not at all, Less than once a week, At least once a week	Categorical	7(0.00%)
13.	Husbands/partners occupation	Job of a partner	Not working, Agricultural-employee, Non-Agriculture-employee	Categorical	131(2%)
14.	Mother's occupation	Respondents occupation	Not working, Agricultural-employee, Non-Agriculture-employee	Categorical	82(1%)
15.	Antenatal care	Maternal healthcare service received or Not during pregnancy from Doc./Nurse/Midwife.	(1)Yes or (0)No	Categorical	21(0.0%)
16.	Assistance during delivery	Maternal healthcare service received or Not during delivery from Doc./Nurse/Midwife.	(1)Yes or (0)No	Categorical	6(0.0%)
17.	Postnatal care	Maternal healthcare service received or Not after delivery within 2 days (48 hrs.) from Doc./Nurse/Midwife.	(1)Yes or (0)No	Categorical	24(0.0%)

4.4 Statistical Summary of the Selected Attributes

The summary of each of the selected attributes used for model building are statistically described in detail in Table 4.2. This statistical summary of the attributes is helpful for understanding of the data set for DM model building phase.

Table 4.2: Statistical summary of the variables

No.	Variables	Frequency	Percent (%)
1.	**Respondent's Age**		
	15-19	416	5.4
	20-24	1596	20.6
	25-29	2292	29.5
	30-34	1507	19.4
	35-39	1203	15.5
	40-44	550	7.1
	45-49	200	2.6
	Missing	0	0
	Total	7,764	100.0
2.	**Place of residence**		
	Rural	6251	80.5
	Urban	1513	19.5
	Missing	0	0
	Total	7,764	100.0
3.	**Region**		
	Tigray	847	10.9
	Afar	714	9.2
	Amhara	965	12.4
	Oromiya	1100	14.2
	Somali	559	7.2
	Benishangul – Gumuz	674	8.7
	SNNP	1053	13.6
	Gambela	608	7.8
	Harari	440	5.7
	Addis Ababa	348	4.5
	Dire Dawa	456	5.9
	Missing	0	0
	Total	7,764	100.0
4.	**Women's highest educational level**		
	No education	5184	66.8
	Primary	2095	27.0
	Secondary	312	4.0
	Higher	173	2.2
	Missing	0	0.0
	Total	7,764	100.0
5.	**Wealth Index**		
	Lowest	2279	29.4
	Second	1354	17.4
	Middle	1241	16.0
	Fourth	1229	15.8
	Highest	1661	21.4
	Missing	0	0.0
	Total	7,764	100.0
6.	**Religion**		
	Orthodox	2694	34.7

	Catholic	79	1.0
	Protestant	1479	19.0
	Muslim	3359	43.3
	Traditional	60	0.8
	Other	89	1.1
	Missing	4	0.1
	Total	7,764	100.0
7.	**Marital status**		
	Never married	66	0.9
	Married	6624	85.3
	Living together	419	5.4
	Widowed	164	2.1
	Divorced	336	4.3
	Not living together	155	2.0
	Missing	0	0
	Total	7,764	100.0
8.	**Husbands educational level**		
	No education	3847	49.5
	Primary	2790	35.9
	Secondary	594	7.7
	Higher	372	4.8
	Missing	161	2.1
	Total	7,764	100.0
9.	**Total children ever born**		
	1 child	1477	19.0
	2 or 3 child	2419	31.2
	4 or 5 child	1778	22.9
	More than 6 child	2090	26.9
	Missing	0	0
	Total	7,764	100.0
10.	**Frequency of reading newspaper**		
	Not at all	7032	90.6
	Less than once a week	576	7.4
	At least once a week	151	1.9
	Missing	5	0.1
	Total	7,764	100.0
11.	**Frequency of listening to radio**		
	Not at all	4280	55.1
	Less than once a week	2189	28.2
	At least once a week	1290	16.6
	Missing	5	0.1
	Total	7,764	100.0
12.	**Frequency of watching television**		
	Not at all	5232	67.4
	Less than once a week	1537	19.8
	At least once a week	988	12.7
	Missing	7	0.1
	Total	7,764	100.0
13.	**Husbands/partners occupation**		

	Not working	148	1.9
	Agriculture –employee	5365	69.1
	Non agriculture – employee	2120	27.3
	Missing	131	1.7
	Total	7,764	100.0
14.	**Respondents occupation**		
	Not working	4035	52.0
	Agriculture –employee	1625	20.9
	Non agriculture – employee	2022	26.0
	Missing	82	1.1
	Total	7,764	100.0
15.	**Antenatal care**		
	0(No)	4828	62.2
	1(Yes)	2915	37.5
	Missing	21	0.3
	Total	7,764	100.0
16.	**Assistance during delivery**		
	0(No)	6485	83.5
	1(Yes)	1273	16.4
	Missing	6	0.1
	Total	7,764	100.0
17.	**Postnatal care**		
	0(No)	6675	86.0
	1(Yes)	1065	13.7
	Missing	24	0.3
	Total	7,764	100.0

4.5 Data Pre-Processing

Much of the raw data contained in databases is unpre-processed, incomplete, and noisy.

For example, the databases may contain:

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

To be useful for data mining purposes, the databases need to undergo pre-processing, in the form of data cleaning and data transformation. Data mining often deals with data that hasn't been looked at for years, so that much of the data contains field values that have expired, are no longer relevant, or are simply missing [69].

4.6 Handling Missing Values

Missing data is a problem that continues to plague data analysis methods. Even as our analysis methods gain sophistication, we continue to encounter missing values in fields, especially in databases with a large number of fields. The absence of information is rarely

beneficial [69]. All things being equal, more data is almost always better. Therefore, we should think carefully about how we handle the thorny issue of missing data [69].

A common method of handling missing values is simply to omit from the analysis the records or fields with missing values. However, this may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting records with missing values would lead to a biased subset of the data. Further, it seems like a waste to omit the information in all the other fields, just because one field value is missing [69]. Therefore, data analysts have turned to methods that would replace the missing value with a value substituted according to various criteria.

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

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

Table 4.3: Attributes with missing values replaced by mode

No.	Attribute	Frequency of missing value	Percentage of missing values	Replaced value
1.	Religion	4	0.1	3(Muslim)
2.	Husbands educational level	161	2.1	1(Not at all)
3.	Frequency of reading newspaper	5	0.1	0(Not at all)
4.	Frequency of listening radio	5	0.1	0(Not at all)
5.	Frequency of watching television	7	0.1	0(Not at all)
6.	Husbands occupation	131	1.7	1 (Agriculture – employee)
7.	Respondents occupation	82	1.1	0 (Not working)
8.	Antenatal care	21	0.3	0(No)
9.	Delivery care	6	0.1	0(No)
10.	Postnatal care	24	0.3	0(No)

4.7 Data Transformation and Reduction

The data may also need to be transformed into forms appropriate for mining. The process of data transformation might include smoothing (e.g. using bin means to replace data errors), Normalization, where the attribute data are scaled so as to fall within a small specified range (scaling the data inside a fixed range), and Attribute construction, where new attributes are constructed and added from the given set of attributes to help the mining process [68].

The data is needed to be reduced in order to make the analysis process manageable and cost-efficient. Data reduction techniques include a data discretization technique which is used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values, data cube aggregation, dimension reduction (irrelevant or redundant attributes are removed), and data compression (data is encoded to reduce the size, numerously reduction (models or samples are used instead of the actual data) [68].

From the original dataset the age attribute was available in two formats: Discrete and Continuous value. Thus for this study the grouped age attribute is selected as shown below:

Table 4.4: A discretized age attribute

Age	Represented value
15-19	1
20-24	2
25-29	3
30-34	4
35-39	5
40-44	6
45-49	7

The attribute number of living children is a continuous variable which discretization was performed to convert into four distinct values as shown in table below:

Table 4.5: A discretized number of living children

Number of total children	Represented value
1	One child
2 – 3	Two or Three children

4 – 5	Four or Five children
6+	More than six children

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

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

Finally, after pre-processing the original dataset assumed to be relevant to the target variable, which consists of 15 variables (12 Predictor (Independent) and 3 Outcome (Dependent)) and 7,764 instances, was used for constructing the model.

4.8 Data Preparation for Weka Software

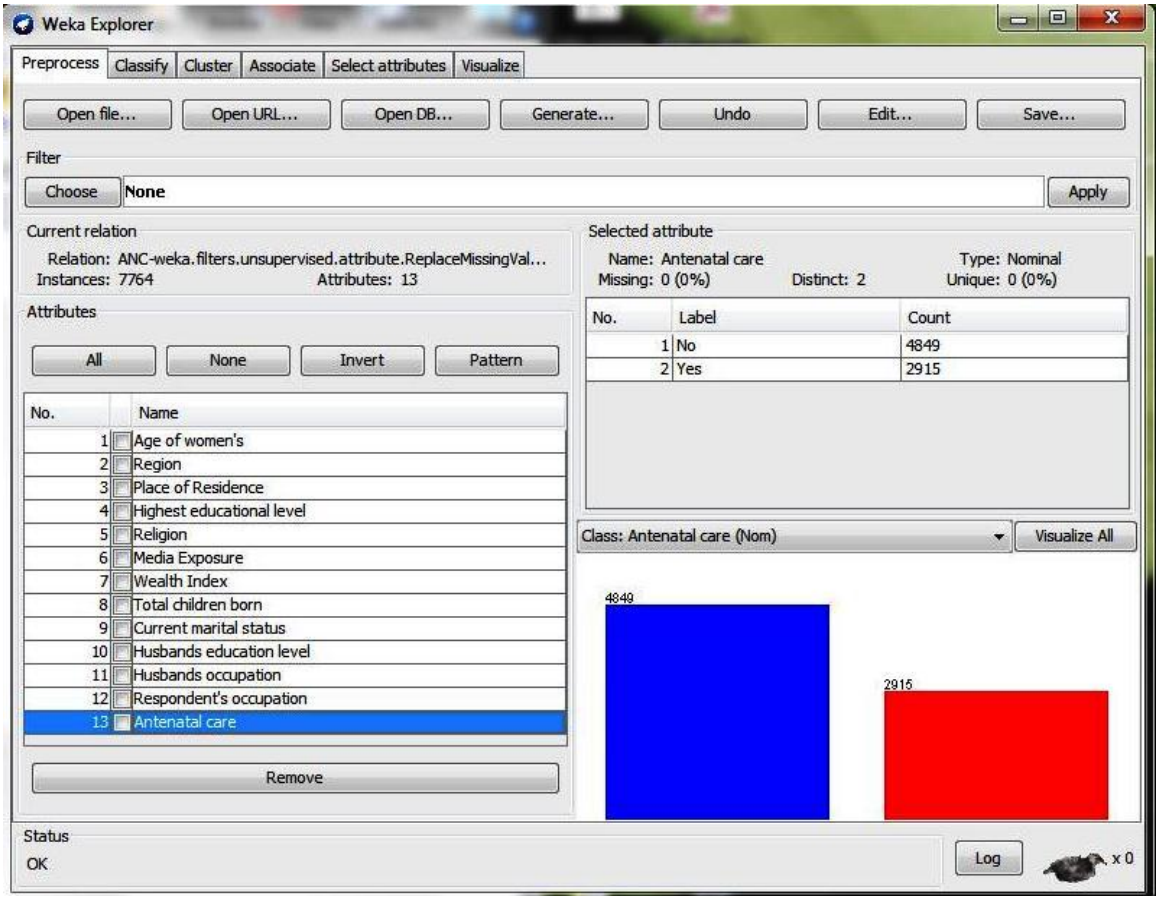
Weka needs the data set to be prepared in some Weka understandable formats. The researcher first has exported the original SPSS file format into Microsoft Excel. Then pre-processing activities are performed and the file is saved into Weka acceptable comma separated values (CSV) or comma delimited file format. Weka native data format is known as the ARFF (Attribute Relation File Format). It is basically a CSV (comma separated value) format with some extra headers to specify what type each attribute is (numerical, binary, nominal). The CSV file format is converted into ARFF by using Weka mining software, to take advantage of easier data manipulation and also compatible interaction with Weka software. During scan of the pre-processed data some basic statistics summary will be produced for each attributes. For categorical attributes, the frequency for each attribute value is shown. By default the last attribute in the list of the attributes name in the dataset is the class (dependent) attribute.

4.9 Setting the Class Attribute

In supervised classification technique predefined classes are required in order to train and test classification models. The setting of predefined class is done intentionally because the employed technique for this study is J48 decision tree and Naïve Bayes classifications. In order to classify women's according to their background variables into their respective classes, the target attribute selected in this research is maternal health care seeking behaviour which consists of three outcome (dependent) variables. These variables are Antenatal, Delivery and Postnatal care each with Binary value (Yes or No). If a woman has received any of the three services from a Doctor/Nurse/Midwife, it is coded as 1(Yes) otherwise 0(N0) is

coded for each of the respondent. Therefore the employed technique was selected in order to classify women's of reproductive age to their respective class based on socio-economic and demographic characteristics.

Figure 4.1: Weka 3.6.8 explorer window with a list of selected attributes



As can be shown in figure 4.1, the two classes consists 7764 cases with 2915 (37.5%) Yes and 4849 (62.5%) No dataset were applied for experimentation for Antenatal care model building.

For Delivery care, the two classes consists 7764 cases with 1273 (16.4%) Yes and 6491 (83.6%) No dataset were applied for experimentation.

Finally for Postnatal care the two classes consists 7764 cases with 1065 (13.7%) Yes and 6699 (86.3%) No dataset were also applied for experimentation to construct the model.

CHAPTER FIVE

EXPERIMENTATION AND ANALYSIS

In the hybrid KDP methodology selected for this study, model building is an iterative process. Therefore, it is important to conduct different experiments to find the best model for solving the problem. Since the main objective of this study was to predict maternal healthcare seeking pattern of women of reproductive age, data mining classification techniques were applied to develop predictive models. In general twelve (12) experiments were conducted using the algorithms for ANC, Delivery and Postnatal care: J48 Decision Tree and Naïve Bayes using Weka 3.6.8 DM software. Moreover, different experiments were conducted in each of the outcome variables and evaluated their performance with the output from the algorithms.

5.1 Experimental Design

Different experiments were constructed for each classifier using the entire dataset consisting of 7,764 instances and 15 attributes including the three outcome (dependent) variables. A stratified 10-fold cross-validation was used to estimate the performance of each classifier. This performance estimation approach has been proved to be statistically good enough in evaluating the performance of data mining classifier algorithms. Overall classification Accuracy, TP rate, FP rate, Size of the tree, Number of leaves and ROC area are used to evaluate and compare the performance of the models generated. These measures were driven from the confusion matrix of the models.

The algorithms used for predictive model building are found in Weka 3.6.8. This version works on many file formats than its predecessors and it is compatible with CSV file format. Thus, no additional effort was exerted to change the dataset from excel to ".arff" file format which is available to save in Weka format. The preprocessed dataset is saved using CSV file format from Excel directly. Then, this file is imported to Weka by clicking on "open" button of explorer window and browsing to the files directory.

5.1.1 Model Building using Decision Tree (J48 algorithm)

J48 is Weka's implementation of the C4.5 algorithm which can work on multiple valued attributes. It contains some parameters that can be changed to improve its performance on different values. Different experiments were conducted for the J48 classifier by changing the main parameters of the algorithm to build a better predictive model.

Table 5.1: Synopsis of the selected J48 classifier parameters

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

BinarySplits parameter by default is set to "False". If this value is changed to "True", it enforces the model generated to be binary decision tree rather than generalized decision tree. The confidence factor helps to set a limit so that the algorithm makes more or less pruning. The default value for confidence factor is 0.25. The minimum number of instances per leaf is set 2 and the last parameters whether pruning is performed or not is defaulted to False mainly affect the performance of the model built.

Moreover, it is to see the effect of tree pruning method on classification accuracy of J48, the tree size, number of leaves and the time taken to build the model by the algorithm. In addition the 10-fold cross validation test option is employed. Since this test option is selected for all experiments due to its performance to train and test the classification model as discussed in Chapter two.

For J48 algorithm four scenarios were considered, with and without pruning for each of the outcome variables.

Scenario 1: J48 decision tree unpruned with all attributes
Scenario 2: J48 decision tree pruned with all attributes
Scenario 3: J48 decision tree unpruned with reduced attributes
Scenario 4: J48 decision tree pruned with reduced attributes

The following experiments was designed to build a model by changing the values of parameter as shown in Table 5.2 for each of the outcome variables used for this study.

Table 5.2: Values of parameters used in the experiments

Experiments	Parameters		
	Unpruned	**Confidence Factor**	**Numbers of Instance(minNumObj)**
Exp#1	True	0.25	2
Exp#2	True	0.30	5
Exp#3	True	0.50	10
Exp#4	False	0.25	2
Exp#5	False	0.15	5
Exp#6	False	0.18	5

5.2.1. J48 Experiment for the outcome variable Antenatal care

Experiment I

The first category of experiment was designed to build the model using J48 algorithm by using all attributes as shown in Table 5.3.

After all necessary J48 decision tree algorithm parameters were set as shown on Table 5.2, the experiment listed in Table 5.3 was conducted. The summary of the result of the six (6) experiments is presented in Table 5.3.

Table 5.3: J48 classifier output using all attribute for ANC

	Experiments					
Performance measure	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Accuracy (%)	70.1%	71.4%	73%	74.4%	74.76%	**74.8%**
WTPR	0.70	0.71	0.73	0.74	**0.75**	**0.75**
WFPR	0.38	0.37	**0.36**	0.37	0.37	0.37
WROC	0.68	0.73	**0.75**	0.71	0.72	0.72
NL	2034	890	494	164	**45**	**45**
ST	2655	1155	632	217	**57**	**57**
Time(sec.)	0.26	0.05	0.05	0.12	0.06	0.05
CCI	5443	5544	5664	5777	5804	**5806**
ICI	2321	2220	2100	1987	1960	**1958**

Key: WTPR-Weighted True Positive rate, WFPR- Weighted False Positive Rate, WROC-Weighted Receiver Operating Characteristics, NL- Number of leaves, ST- Size of the tree, CCI-Correctly Classified Instances, ICI- Incorrectly Classified Instances.

The model built from Exp #1-3 with unpruned parameter depicts that it increases its accuracy by 1.5 whereas for pruned parameter from Exp #4-6 it shows almost similar accuracy. The same is true in WTPR it increases for unpruned and almost show similar result for pruned parameter. WROC get its maximum for unpruned parameter conducted in Exp #3.

The overall performance of the models generated shows that maximum performance was achieved when the value of the confidence factor and number of instance per leaf kept at 0.18 and 5 respectively. The model built on Exp #6 with pruned parameter gives maximum performance for the first category of experiment. Therefore, the last experiment is good based on accuracy, the number of leaves and the size of the tree produced than the other experiments. The tree generated for this experiment is shown on Appendix A.

Experiment II

The second category of experiment was designed to build the model with J48 algorithm with a reduced attributes.

Before conducting these experiments attributes selection is necessary. This is because of the fact that irrelevant attributes may lead to poor decision tree model. Since attribute selection is important in decision tree models, the researcher ranked the attributes based on information gain, Information gain measures the orders of the attributes computed using the formula. Ranking the attributes to the mining task of the decision tree was implemented by Weka attribute ranking filter versus information gain.

Table 5.4: List of selected attributes with their information gain

No.	Ranked attributes	Information Gain
1	7 Wealth Index	0.149
2	3 Place of Residence	0.123
3	4 Highest educational level	0.097
4	11 Husbands occupation	0.096
5	2 Region	0.079
6	10 Husbands education level	0.075
7	6 Media Exposure	0.057
8	8 Total children born	0.031

As shown in Table 5.2 the parameters are used to conduct the six (6) experiments to see the effect of attribute selection on ANC as shown in Table 5.5. The summary of the classifier output of the experiments is presented in Table 5.5:

Table 5.5: J48 classifier output using selected attribute for ANC

	Experiments					
Performance	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy	71.5%	72..4%	73.8%	74.5%	74.4%	**74.5%**
WTPR	0.72	0.72	0.74	**0.75**	0.74	0.74
WFPR	0.38	0.37	**0.36**	0.38	0.38	0.37
WROC	0.72	0.74	**0.76**	0.72	0.72	0.72
NL	1022	513	274	93	**28**	**28**
ST	1349	667	344	119	**38**	**38**
Time(sec.)	0.04	0.04	0.03	0.03	0.03	0.03
CCI	5553	5624	5727	5782	5773	**5785**
ICI	2211	2140	2037	1982	1991	**1979**

As can be seen from the Table 5.5 the experiment reveals that for unpruned parameter as the confidence factor and number of instance per leaf increases its accuracy increases by 0.5. WTPR and WROC also give maximum result in Exp #3. Whereas for pruned parameter even though these parameters were changed it shows almost similar output through Exp #4-5. These three parameters are used for optimizing accuracy and size of decision tree (pruning tactics) in testing set of data for model building.

Finally the model built with pruned parameter in Exp #6 shows good performance than the other experiments conducted in the second category for reduced attributes.

5.2.2. J48 Experiment on the outcome variable Delivery care

This experiment was conducted on the outcome variable Assistance during delivery using all and reduced attributes to build the model based on the parameter set in Table 5.2.

Experiment III

The third category of experiment was conducted to build the model using the J48 algorithm by using all attributes as shown in Table 5.6.

Table 5.6: The J48 classifier output with all attributes for Assistance care

	Experiments					
Performance	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy	89.7%	90.4%	91.23%	**91.35%**	91.12%	91.1%
WTPR	0.89	0.90	0.91	0.91	0.91	0.91
WFPR	0.33	0.32	0.30	0.29	0.29	0.30
WROC	0.81	0.86	**0.88**	0.83	0.84	0.84
NL	851	451	204	97	19	29
ST	1055	552	253	126	24	38
Time(sec.)	0.21	0.03	0.02	0.04	0.04	0.04
CCI	6965	7021	7083	7093	7075	7074
ICI	799	743	681	671	689	690

From the Table 5.6 the classifiers performance for unpruned parameter shows that the accuracy, WTPR and WROC reach the highest performance when the confidence factor and minimum number of instance per leaf is 0.5 and 10, respectively. Whereas for pruned parameter conducted in Exp #4-6 the classifiers performance for accuracy decreases, whereas WTPR and WROC show similar result. The accuracy is highest when the confidence factor and minimum number of instance per leaf is set 0.25 and 2, respectively. Over all the third experiment (Exp #3) shows a better performance with respect to WROC than the other experiments conducted under this category.

Experiment IV

The fourth category of experiment was conducted to investigate the effect of reduced attributes on the performance of J48 algorithm with all attributes conducted in Table 5.6.

Before conducting the experiment attribute selection was performed on Weka attribute ranking filter and information gain from selectAttributes command.

Table 5.7: List of selected attributes with their information gain

No.	**Ranked attributes**	**Information Gain**
1	3 Place of Residence	0.227
2	6 Wealth Index	0.222
3	9 Husbands Occupation	0.16

4	4 Highest educational level	0.139
5	2 Region	0.13
6	8 Husbands education level	0.11
7	5 Media Exposure	0.095
8	7 Total children born	0.061
9	1 Age of women's	0.009

By using select Attribute command from Weka's explorer window, 9 attributes from the total of 12 predictor variables were selected according to descending order of information gain as shown in Table 5.7. The summary of the J48 classifier output for the six experiments is shown in Table 5.8.

Table 5.8: The J48 classifier output on reduced attributes

	Experiments					
Performance	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy	90.2%	90.7%	**91.2%**	**91.2%**	91.0%	91.0%
WTPR	0.90	0.91	**0.92**	0.91	0.91	0.91
WFPR	0.31	0.30	0.29	0.29	0.28	0.28
WROC	0.84	0.87	**0.89**	0.83	0.84	0.84
NL	631	297	129	79	23	23
ST	796	367	175	105	28	28
Time(sec.)	0.05	0.04	0.03	0.03	0.03	0.03
CCI	7004	7044	7084	7085	7069	7068
ICI	760	720	680	679	695	696

From the Table 5.8 shown the classifiers performance for unpruned parameter conducted in Exp #1-3 increases its accuracy, WTPR, and WROC. The classifier achieved the highest performance in Exp #3 when the confidence factor and minimum number of instance per leaf is 0.5 and 10, respectively. Whereas for pruned parameter the classifiers performance for accuracy, WTPR and WROC decreases and reach the highest value when the confidence factor and minimum number of instance per leaf is set 0.25 and 2, respectively. Over all the experiments conducted in Exp #3 and #4 shows a better performance than the others under this category of experiment. However, the classifier output conducted with unpruned parameter in Exp #3 is better than Exp #4 with respect to WROC. The tree generated for this experiment is shown on Appendix B.

5.2.3. J48 Experiment for the outcome variable Postnatal care

This experiment was performed to build the model using J48 algorithm for the outcome variable postnatal care using all and selected attributes in experiment V and VI respectively.

Experiment V

The fifth category of experiment was conducted to build the model using J48 algorithm with all attributes based on the value of parameter set on Table 5.2. The summary of the classifier outputs of the experiments is presented in Table 5.9.

Table 5.9: The J48 classifier output with all attributes for Postnatal care

	Experiments					
Performance	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy	85.75%	86.6%	87.3%	87.9%	87.9%	88%
WTPR	0.86	0.87	0.87	0.88	0.88	0.88
WFPR	0.57	0.59	0.56	0.58	0.57	0.58
WROC	0.69	0.77	0.79	0.75	0.75	0.75
NL	1172	498	199	88	35	27
ST	1482	620	244	115	46	34
Time(sec.)	0.06	0.05	0.05	0.06	0.05	0.05
CCI	6658	6720	6779	6827	6828	6833
ICI	1106	1044	985	937	936	931

From the Table 5.9 the classifiers performance for unpruned parameter shows the accuracy, WTPR, and WROC reach the highest when the confidence factor and minimum number of instance per leaf is 0.5 and 10, respectively. On the other hand for pruned parameter the classifiers performance for accuracy and WTPR reach the highest value when the confidence factor and minimum number of instance per leaf is 0.18 and 5, respectively. But for WROC the result for Exp #3 is greater than others. Over all the Exp #3 shows a better performance than the others with respect to accuracy, WTPR and WROC.

Experiment VI

The sixth category of experiment was conducted using reduced attribute to investigate the effect of classifier performance using J48 algorithm on all attributes conducted in Table 5.9.

According to the experiment conducted with reduced attribute for J48 algorithm for ANC and Delivery care, attribute selection was run on Weka using information gain and ranking attribute command to select the best attributes. Hence, the following 8 attributes were selected as follows according to their Information gain value as shown in Table 5.10.

Table 5.10: List of selected attributes with their information gain

No.	Ranked attributes	Information Gain
1	7 Wealth Index	0.109
2	3 Place of Residence	0.108
3	11 Husbands occupation	0.084
4	4 Highest educational level	0.082
5	10 Husbands educational level	0.068
6	2 Region	0.062
7	6 Media Exposure	0.054
8	7 Total children born	0.028

Therefore, after selecting the 8 attributes as shown on Table 5.10 the summary of the J48 algorithm classifier output is presented in Table 5.11.

Table 5.11: The J48 classifier output with reduced attributes for Postnatal care

	Experiments					
Performance	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy	86.3%	87.1%	87.5%	88.5%	88.3%	88.1%
WTPR	0.87	0.87	0.88	0.89	0.88	0.88
WFPR	0.57	0.58	0.58	0.54	0.56	0.57
WROC	0.73	0.79	0.80	0.75	0.75	0.75
NL	667	373	184	48	34	28
ST	848	465	223	64	44	35
Time(sec.)	0.04	0.03	0.02	0.03	0.03	0.02
CCI	6726	6759	6797	6870	6857	6843
ICI	1038	1005	967	894	907	921

From the Table 5.11 as shown for unpruned parameter the accuracy, WTPR and WROC increases from Exp #1-3 and achieves the highest performance on Exp #3 with 87.5%, 88%, and 0.80, respectively. For pruned parameter from Exp #4-6 the accuracy is decreasing,

whereas WTPR and WROC show almost similar performance result. In general, the model built in Exp #3 with unpruned parameter shows a better performance than the others under this category of experiment. The tree generated for this experiment is shown on Appendix C.

5.1.2 Model building using Naïve Bayes algorithm

The second type of classification technique applied in this study is the Naïve Bayes algorithm. Six (6) experiments were conducted using all and reduced attributes for each of the three outcome variables. In the experiment the 10-fold cross validation was used for all experiments. The Naïve Bayes experiment was designed to build the model for predicting maternal health care seeking pattern and to compare the performance with J48 algorithm. In this experiment two scenarios were considered, one containing all 12 predictors and the other containing the selected 8 attributes for each of the outcome variables (ANC, Delivery and Postnatal care).

Experiment VII

This experiment was conducted on the outcome variable ANC using all attributes and selected attributes to build the model.

Table 5.12: Naïve Bayes classifier output for ANC

Model	Accuracy	WTPR	WFPR	WROC
Naïve Bayes with all attributes	74.77%	0.75	0.34	0.78
Naïve Bayes with selected attributes	74.72%	0.75	0.34	0.77

The Naïve Bayes model built on all attributes correctly classified 5805(74.77%) instances while 1959(25.23%) instances were classified incorrectly. The second model built on selected attributes correctly classified 5751(74.72%) instances while 2013(25.3%) were classified incorrectly.

Experiment VIII

The experiment was conducted on the outcome variable Delivery care using all attributes and selected attributes to build the model. The summary of the classifier output is shown in Table 5.13.

Table 5.13: Naïve Bayes classifier output for Delivery care

Model	Accuracy	WTPR	WFPR	WROC
Naïve Bayes with all attributes	90.0%	0.9	0.2	0.91
Naïve Bayes with selected attributes	89.7%	0.89	0.205	0.90

The Naïve Bayes model built on all attributes correctly classified 6989(90.0%) instances while 755(9.98%) instances were classified incorrectly. The second model built on selected attributes correctly classified 6954(89.56%) instances while 810(10.43%) were classified incorrectly.

Experiment IX

This experiment was conducted on the outcome variable Postnatal care using all attributes and selected attributes to build a model using Naïve Bayes algorithm. The summary of the classifier output is shown in Table 5.14.

Table 5.14: Naïve Bayes classifier output for Postnatal care

Model	Accuracy	WTPR	WFPR	WROC
Naïve Bayes with all attributes	85.5%	0.86	0.32	0.83
Naïve Bayes with selected attributes	85.4%	0.85	0.32	0.83

The Naïve Bayes model built on all attributes correctly classified 6639(85.5%) instances while 1125(14.5%) instances were classified incorrectly. The second model built on selected attributes correctly classified 6602(85.4%) instances while 1162(14.6%) instances were classified incorrectly.

5.1.3 Evaluating the Discovered Knowledge

After developing the models using J48 decision tree and Naïve Bayes algorithms the model evaluation step are applied to select the best model for each of the outcome variable. For the experiments conducted for J48 classification algorithm accuracy, WTPR, WFPR, WROC, number of leaves and size of the tree are considered to evaluate the performance of the model generated in six experiments (I-VI) whereas for experiments conducted for Naïve Bayes algorithm accuracy, WTPR, WFPR and WROC are used for model evaluation phase of data mining Hybrid methodology as discussed in section 3.5. Model evaluation was conducted for the three outcome variables as follows:

5.1.3.1 Model Evaluation for the Antenatal Care

First the experiment conducted using J48 decision tree algorithm with all and selected attributes were compared to select the best model as shown in Table 5.15.

Table 5.15: J48 performance evaluation for ANC

Model	Experiment #	Accuracy	WTPR	WFPR	WROC	NL	ST	Time(sec.)
J48 with all attributes	#1	70.1%	0.70	0.38	0.68	2034	2655	0.26
	#2	71.4%	0.71	0.37	0.73	890	1155	0.05
	#3	73%	0.73	0.36	0.75	494	632	0.05
	#4	74.4%	0.74	0.37	0.71	164	217	0.12
	#5	74.76%	0.75	0.37	0.72	45	57	0.06
	#6	**74.8%**	**0.75**	0.37	0.72	45	57	0.05
J48 with reduced attributes	#1	71.5%	0.72	0.38	0.72	1022	1349	0.04
	#2	72..4%	0.72	0.37	0.74	513	667	0.04
	#3	73.8%	0.74	**0.36**	0.76	274	344	0.03
	#4	74.5%	**0.75**	0.38	0.72	93	119	0.03
	#5	74.4%	0.74	0.38	0.72	28	38	0.03
	#6	**74.5%**	0.74	0.37	0.72	28	38	0.03

The model built on the first category of experiment with pruned parameter was selected as best performing model using J48 with all attributes as shown in Exp #6.

Then this model is also compared with Naïve Bayes algorithm based up on the data mining evaluation metrics as shown in Table 5.16.

Table 5.16: J48 and Naïve Bayes algorithm performance evaluation for ANC

Model	Accuracy	WTPR	WFPR	WROC
J48 algorithm with all attributes	**74.8%**	0.75	0.37	0.72
Naïve Bayes with all attributes	74.7%	0.75	0.34	**0.78**
Naïve Bayes with selected attributes	74.7%	0.75	0.34	0.77

The experiment conducted with J48 algorithm with all attributes outperforms Naïve Bayes algorithm as shown in Table 5.16 since it correctly classified 5806 (74.8%) instances while 1958 (25.2%) of the instance were classified incorrectly out of 7,764 instances.

The Confusion Matrix of the model shows the number of instances of each class that are assigned to all possible classes according to the classifier's prediction. The columns represent the predictions, and the rows represent the actual class as depicted in Table 5.17.

Table 5.17: Confusion Matrix of the J48 model for ANC

Actual class	**Predicted class**		Total
	No	Yes	
No	4511	338	4849
Yes	1620	1295	2915
Total	6131	1633	7764

From the Table 5.17 the confusion matrix shows that 4511 instances were correctly predicted as not received Antenatal care (True positive). The number of instance which was correctly predicted as Yes (ANC received) was 1295 instances (True negative). Therefore, correctly classified instances are the sum of these two numbers (sum of diagonal values of the table) which is 5806 and from this its accuracy is 5806/7764 which is 74.78%.

The following result has been extracted from the model. True Positive rate shows the percentage of Yes instances whose predicted values of the class attribute are identical with the actual class. FP rate shows the percentage of instances whose predicted values of the class attribute are not identical with the actual values.

Table 5.18: Summary of the J48 classifier result of ANC

	=== Detailed Accuracy by Class ===						
	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
	0.93	0.56	0.74	0.93	0.82	0.72	No
	0.44	0.07	0.79	0.44	0.57	0.72	Yes
Weighted Avg.	0.75	0.37	0.76	0.75	0.73	0.72	

5.1.3.2 Model Evaluation for Delivery Care

The model evaluation was carried out first for J48 algorithm for all and reduced attributes conducted in Table 5.6 and 5.8, respectively.

Table 5.19: J48 performance evaluation for Delivery assistance care

Model	Experiment #	Accuracy	WTPR	WFPR	WROC	NL	ST	Time(sec.)
J48 with all attributes	#1	89.7%	0.89	0.33	0.81	851	1055	0.21
	#2	90.4%	0.90	0.32	0.86	451	552	0.03
	#3	91.23%	0.91	0.30	0.88	204	253	0.02
	#4	91.35%	0.91	0.29	0.83	97	126	0.04
	#5	91.12%	0.91	0.29	0.84	19	24	0.04
	#6	91.1%	0.91	0.29	0.84	29	38	0.04
J48 with reduced attributes	#1	90.2%	0.90	0.31	0.84	631	796	0.05
	#2	90.7%	0.91	0.30	0.87	297	367	0.04
	#3	**91.2%**	**0.92**	0.29	**0.89**	**129**	**175**	**0.03**
	#4	91.2%	0.91	0.29	0.83	79	105	0.03
	#5	91.0%	0.91	**0.28**	0.84	23	28	0.03
	#6	91.0%	0.91	**0.28**	0.84	23	28	0.03

As can be shown in Table 5.19 the model built using reduced attributes conducted on Exp #3 is selected as the best model with an accuracy of 91.2%, WTPR (0.92) and WROC (0.89).

In addition, the model built using J48 algorithm is compared with Naïve Bayes to select the best algorithm to build the model as shown in Table 5.20.

Table 5.20: J48 and Naïve Bayes algorithm performance evaluation for Delivery care

Model	Accuracy	WTPR	WFPR	WROC
J48 algorithm with reduced attributes	**91.2%**	**0.92**	0.29	0.89
Naïve Bayes with all attributes	90.0%	0.9	**0.2**	**0.91**
Naïve Bayes with selected attributes	89.7%	0.89	0.205	0.90

Therefore, the model built with J48 algorithm outperforms Naïve Bayes with an accuracy of 91.2%, WTPR (0.92) and WROC (0.89) as shown in Table 5.20. The J48 classifier output for this experiment was used to predict Assistance during delivery as shown in Table 5.21.

The confusion matrix of the model is discussed in Table 5.21.

Table 5.21: Confusion Matrix for Delivery care service

Actual Class	**Predicted Class**		Total
	No	Yes	
No	6254	237	6491
Yes	444	829	1273
Total	6698	1066	7764

As indicated in Table 5.21 from the total test dataset (7,764 instances) given to the model, the accuracy rate is (TP + TN) / (TP+TN+FP+FN). The accuracy was 91.23% with 7083 instances are correctly classified while the remaining 681 (8.7%) instances were classified incorrectly.

The following result has been extracted from the model to show the classifier performance as shown in Table 5.22.

Table 5.22: Summary of the J48 classifier result for Delivery care

	=== Detailed Accuracy by Class ===						
	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
	0.963	0.349	0.934	0.964	0.948	0.890	No
	0.651	0.037	0.778	0.651	0.713	0.890	Yes
Weighted Avg.	0.912	0.298	0.91	0.912	0.91	0.890	

5.1.3.3 Model Evaluation for Postnatal Care

The same techniques was applied for this outcome variable in order to evaluate the models developed using J48 and Naïve Bayes classification algorithm with all and reduced attributes.

First the model built using J48 algorithm with all and reduced attributes were compared to select the best model conducted in Table 5.9 & 5.11 as revealed in Table 5.23.

Table 5.23: J48 performance evaluation for Postnatal care

Model	Experiment #	Accuracy	WTPR	WFPR	WROC	NL	ST	Time(sec.)
J48 with all attributes	#1	85.75%	0.86	0.57	0.69	1172	1482	0.06
	#2	86.6%	0.87	0.59	0.77	498	620	0.05
	#3	87.3%	0.87	0.56	0.79	199	244	0.05
	#4	87.9%	0.88	0.58	0.75	88	115	0.06
	#5	87.9%	0.88	0.57	0.75	35	46	0.05
	#6	88%	0.88	0.58	0.75	27	34	0.05
J48 with reduced attributes	#1	86.3%	0.87	0.57	0.73	667	848	0.04
	#2	87.1%	0.87	0.58	0.79	373	465	0.03
	#3	87.5%	0.88	0.58	**0.80**	**184**	**223**	**0.02**
	#4	**88.5%**	**0.89**	**0.54**	0.75	48	64	0.03
	#5	88.3%	0.88	0.56	0.75	34	44	0.03
	#6	88.1%	0.88	0.57	0.75	28	35	0.02

As shown in Table 5.23 from the experiments conducted using J48 algorithm the model built with reduced attributes conducted in Exp #3 with unpruned parameter was selected as best performing model to predict postnatal care. Thus this model is compared with experiment conducted with Naïve Bayes algorithm as shown in Table 5.24.

Table 5.24: J48 and Naïve Bayes algorithm performance evaluation for Postnatal care

Model	Accuracy	WTPR	WFPR	WROC
J48 algorithm with reduced attributes	**87.5%**	**0.88**	0.58	0.80
Naïve Bayes with all attributes	85.5%	0.86	**0.32**	**0.83**
Naïve Bayes with selected attributes	85.4%	0.85	**0.32**	**0.83**

Accordingly, the best model selected from the two algorithms is the J48 algorithm built with reduced attributes conducted on the first scenario to predict postnatal care having an accuracy of 87.5%, WTPR of (0.88) and WROC (0.80) as shown in Table 5.24.

The confusion matrix of the model is shown in Table 5.25:

Table 5.25: Confusion Matrix for Postnatal care

Actual class	**Predicted class**		**Total**
	No	Yes	
No	6442	257	6699
Yes	710	355	1065
Total	7152	612	7764

As shown in Table 5.25 above it can be observed that from the total test set (7,764 instances) given to the model, the accuracy rate is (TP + TN)/ (TP+TN+FP+FN). The accuracy was 87.5% out of the total 6797 instances are correctly classified while the remaining 967 (12.4%) instances were classified incorrectly.

The following result has been extracted from the model to show the classifier performance as shown in Table 5.26.

Table 5.26: Summary of the J48 classifier output for postnatal care

	=== Detailed Accuracy by Class ===						
	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
	0.962	0.667	0.901	0.962	0.93	0.80	No
	0.333	0.038	0.58	0.333	0.423	0.80	Yes
Weighted Avg.	0.875	0.58	0.857	0.875	0.861	0.80	

5.1.4 Generating Rules from J48 Decision Trees

One of the most attractive aspects of decision trees lies in their interpretability, especially with respect to the construction of decision rules. Decision rules can be constructed from a decision tree simply by traversing any given path from the root node to any leaf. The complete set of decision rules generated by a decision tree is equivalent (for classification purposes) to the decision tree itself [73].

Decision rules come in the form *if antecedent, then consequent*, as shown below for each of the outcome variable. For decision rules, the antecedent consists of the attribute values from the branches taken by the particular path through the tree, while the consequent consists of the classification value for the target variable given by the particular leaf node [73].

The researcher selects and discusses a rule that satisfy the assumptions of the domain expert. Based on this assumption, some of the best rules/patterns are extracted from the decision tree

for each of the outcome variables that contain most instances of the dataset was selected. The following rules indicate that the possible conditions in which a woman could be classified in each of the maternal health care utilization method classes.

5.1.4.1 Rules Generated from J48 Decision Tree for Antenatal Care

Rule1. If place of residence = Rural AND Highest educational level = No education, then Antenatal care not received (4665.0/1065.0).

Rule2. If place of residence = Rural AND Highest educational level = Primary, AND Husbands occupation = Agricultural-employee then Antenatal care Not received (1173.0/436.0).

Rule3. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Non-agricultural-employee AND Religion = Orthodox then Antenatal care received (82.0/25.0).

Rule4. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Non-agricultural-employee AND Religion = Muslim AND current marital status = living together AND wealth index = Richest then Antenatal care received (24.0/7.0).

Rule5. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Non-agricultural-employee AND Religion = Protestant AND Region = Oromiya then Antenatal care received (11.0/3.0).

Rule6. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Non-agricultural-employee AND Religion = Protestant AND Region = SNNP then Antenatal care received (41.0/12.0).

Rule7. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Non-agricultural-employee AND Religion = Protestant AND Region = Gambela then Antenatal care received (49.0/23.0).

Rule8. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Non-agricultural-employee AND Religion = Catholic then Antenatal care not received (8.0/1.0).

Rule9. If place of residence = Rural AND Highest educational level = Primary AND Husbands occupation = Not working, then Antenatal care not received (44.0/14.0).

Rule10. If place of residence = Urban AND Wealth Index = Richer AND Husbands education level = primary, then Antenatal care received (46.0/18.0).

Rule11. If place of residence = Urban AND Wealth Index = Richer AND Husbands education level = secondary, then Antenatal care received (10.0/3.0).

5.1.4.2 Rules Generated from J48 Decision Tree for Delivery Care

Rule1. If place of residence = Rural AND Highest educational level = primary AND current marital status = living with partner AND Husbands occupation = Agricultural-employee AND Total children born = Two or three child AND Religion = Orthodox AND Wealth Index = poorer, then Assistance care not received (27.0/1.0).

Rule2. If place of residence = Rural AND Highest educational level = primary AND current marital status = living with partner AND Husbands occupation = Agricultural-employee AND Total children born = Two or three child AND Religion = Muslim, then Assistance care not received (118.0/3.0).

Rule3. If place of residence = Rural AND Highest educational level = primary AND current marital status = living with partner AND Husbands occupation = Agricultural-employee AND Total children born = Two or three child AND Religion = Protestant, then Assistance care not received (120.0/8.0).

Rule4. If place of residence = Rural AND Highest educational level = primary AND current marital status = living with partner AND Husbands occupation = Agricultural-employee AND Total children born = Six or more child, then Assistance care Not received (230.0/5.0).

Rule5. If place of residence = Rural AND Highest educational level = primary AND current marital status = living with partner AND Husbands occupation = Agricultural-employee AND Total children born = Four or Five child, then Assistance care Not received (218.0/8.0).

Rule6. If place of residence = Urban AND wealth index = Richest AND Media exposure = No media access AND Highest education level = primary AND Region = Addis Ababa, then Assistance care received (49.0/10.0).

Rule7. If place of residence = Urban AND wealth index = Richest AND Media exposure = No media access AND highest education level = secondary, then Assistance care received (48.0/9.0).

Rule8. If place of residence = Urban AND wealth index = Richest AND Media exposure = media access AND Total children born = Two or three child AND Highest educational level

= Primary AND Age of women's = 20-34 AND Region = Addis Ababa, then Assistance care received (36.0/8.0).

Rule9. If place of residence = Urban AND wealth index = Richest AND Media exposure = media access AND Total children born = Two or three child AND Highest educational level = Primary AND Age of women's = 20-34 AND Region = Harar, then Assistance care received (25.0/3.0).

Rule10. If place of residence = Urban AND wealth index = Richest AND Media exposure = media access AND Total children born = Two or three child AND Highest educational level = Primary AND Age of women's = 20-34 AND Region = Dire Dawa, then Assistance care received (18.0/2.0).

Rule11. If place of residence = Urban AND wealth index = Richest AND Media exposure = media access AND Total children born = Two or three child AND Highest educational level = Secondary, then Assistance care received (86.0/6.0).

Rule12. If place of residence = Urban AND wealth index = Richest AND media exposure = media access AND Total children born = One child, then Assistance care received (340.0/44.0).

5.1.4.3 Rules Generated from J48 Decision Tree for Postnatal Care

Rule1. If place of residence = Rural, AND Highest educational level = Primary AND Husbands occupation = Agricultural – employee AND Region = Amhara, then Postnatal care not received (127.0/4.0).

Rule2. If place of residence = Rural, AND Highest educational level = Primary AND Husbands occupation = Agricultural – employee AND Region = Oromiya, then Postnatal care not received (253.0.0/14.0).

Rule3. If place of residence = Rural, AND Highest educational level = Primary AND Husbands occupation = Agricultural – employee AND Region = SNNP, then Postnatal care not received (259.0.0/10.0).

Rule4. If place of residence = Rural, AND Highest educational level = Secondary AND Total children born = Two or Three child, then Postnatal care not received (24.0/4.0).

Rule5. If place of residence = Rural, AND Highest educational level = Secondary AND Total children born = One child AND Husbands educational level = Higher, then Postnatal care received (11.0/5.0).

Rule6. If place of residence = Urban, AND Wealth index = Richest AND Husbands education level = Primary AND Region = Harari AND Total children born = one child, then Postnatal care received (10.0).

Rule7. If place of residence = Urban, AND Wealth index = Richest AND Husbands education level = Secondary AND Region = Addis Ababa AND Total children born = one child, then Postnatal care received (19.0/8.0).

Rule8. If place of residence = Urban, AND Wealth index = Richest AND Husbands education level = Secondary AND Region = Harari, then Postnatal care received (10.0/3.0).

Rule9. If place of residence = Urban, AND Wealth index = Richest AND Husbands education level = Secondary AND Region = Tigray, then Postnatal care received (15.0/5.0).

Rule10. If place of residence = Urban, AND Wealth index = Richest AND Husbands education level = Secondary AND Region = Dire Dawa, then Postnatal care received (29.0/10.0).

Rule11. If place of residence = Urban AND Highest educational level = Secondary AND Region = Addis Ababa, then Postnatal care received (80.0/23.0).

Rule12. If place of residence = Urban AND Highest educational level = Secondary AND Region = Harari, then Postnatal care received (40.0/11.0).

Rule13. If place of residence = Urban AND Highest educational level = Higher AND Region = Addis Ababa AND Total children born = Two or Three child, then Postnatal care received (15.0/5.0).

Rule14. If place of residence = Urban AND Highest educational level = Higher AND Region = Addis Ababa AND Total children born = One child AND Husbands education level = secondary, then Postnatal care received (10.0/2.0).

Rule15. If place of residence = Urban AND Highest educational level = Higher AND Region = Addis Ababa AND Total children born = One child AND Husbands education level = Higher, then Postnatal care received (13.0/3.0).

Discussion on the generated rules from the classification models

From the generated rules it is observed that the most determinant factors are Place of residence, Household wealth index, Women's educational level, Partner's occupation, Partner's educational level, Region, Number of living children, Religion, Media exposure.

A woman who doesn't have any of the educational level has no chance of using maternal health care services (ANC, Delivery care and Postnatal care). Rural women who are educated to primary level and their husbands are non-agricultural employee lives in Oromiya/SNNP/Gambela whose religion is protestant has a high probability of using ANC.

A rural women who are married and educated to primary level and their partner is agricultural employee having two or three children whose religion is Muslim/Orthodox/Protestant with household wealth index is poor have no probability of receiving assistance care from a skilled personnel.

An urban women whose household wealth index is richest who have media access having two or three children educated to primary level aged 20 – 34 living in Addis Ababa/Harar/Dire Dawa have a high probability of receiving assistance care during delivery from a skilled personnel.

A rural women who are educated to primary level and their partner is agricultural employee living in Amhara/Oromiya/SNNP have no chance of receiving postnatal care. Rural women who are educated to secondary level having only one child and their partner educational level is higher have a high probability of receiving postnatal care.

An urban women whose household wealth index is richest and their partner educational level is secondary having only one child and living in Addis Ababa/Harar/Tigray/Dire Dawa have a high probability of receiving postnatal care from a skilled personnel.

Some of the interesting rules show that education is so important, even though women whose husband are non-agricultural employee when their partners are educated to at least up to primary level is have a chance of getting maternal health care service from a skilled personnel.

In general, women's education level, partner's occupation and educational level increases their likelihood to receive maternal health care service from skilled personnel.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 CONCLUSION

The major objective of this research was to predict MHCS pattern from 2011 EDHS datasets using data mining techniques. Different experiments were conducted using J48 and Naïve Bayes classification algorithm with their default and adjusted parameters. In addition, the experiments were conducted using all and selected attributes for J48 and Naïve Bayes algorithms. Relevant attributes were selected for J48 algorithm using Weka's built in InfoGainAttributeEval versus Ranker function for the three outcome variables.

In data mining application, first the data in hand and the business problem to be solved must be analyzed and understood very well. Suitable mining techniques also play an important role for successful data mining application along with data preprocessing. Much emphasis was given for business understanding and data understanding to make sure that the best possible results are obtained. To clean the data in this study, missing values were handled. Moreover, data integration and transformation were managed.

In this research, the methodology employed was Hybrid Data mining process model; it involves six steps and the researcher thoroughly passes through all the steps and iterated as needed. A total of 7,764 women of reproductive age for the last birth during five year preceding the survey conducted were used to build the model. In order to build the models that can predict MHCS in Ethiopia, several experiments were conducted with diverse parameters values. By comparing the overall accuracy, WTPR, WFPR, WROC, number of leaves and size of the tree the best performing algorithm, found in this research, is J48 decision tree than its counterpart Naïve Bayes for each of the three outcome variables (ANC, Delivery and Postnatal care).

The first model selected for the outcome variable ANC with an accuracy rate of 74.8% and WTPR (75%). For the second outcome variable Delivery care the model was selected with an accuracy rate of 91.23%, WTPR (92%), WFPR (29%) and WROC (0.89). The model selected for the outcome variable Postnatal care with an accuracy rate of 87.5%, WTPR (88%), and WROC (0.80).

In general, the results obtained from this study were interesting and encouraging; it can be used as decision support for healthcare practitioner. The extracted rules for each of the

outcome variables are very effective for the prediction of MHCS. From the socio-economic and demographic variables used as predictor variables in this study for each of the outcome variables, it can be observed that the attributes such as Household Wealth Index, Place of Residence, Women's Educational level, Husbands Occupation, Region, Husbands Educational level, Media exposure and Total number of children born are the most determinant factors to predict MHCS.

6.2 RECOMMENDATIONS

Based on the findings obtained from the research, the researcher makes the following recommendation for the healthcare practitioner to support decision making, policies and programs that will improve MHCS through:

- Promoting and advertising using Medias the benefit of receiving maternal health care service in all parts of the country.
- There should be strong monitoring and evaluation system on Maternal and Child Health among FMOH, Regional and Wereda health bureaus.
- Health facilities especially those who are working on maternity health service should keep their records computerized during pregnancy (ANC), during delivery and after delivery (POST) that would be useful to predict MHCS in addition to keeping the records.
- In this study Decision tree and Naïve Bayes data mining techniques are applied to predict MHCS. However, more machine learning algorithms like Artificial Neural Network, Support Vector Machine and Multilayer Perceptron along with much larger data size needs to be taken to recognize the effects and optimize the prediction.
- The decision tree algorithm has achieved interesting results. Hence, an attempt should be made to develop knowledge based system (KBS) that would be helpful in assisting expert advice to identify the actual and non-actual user.

References

1. WHO. Maternal mortality ratio (per 100,000 live births). Available from: http://who.int/healthinfo/statistics/indmaternalmortality/en/index.html; 2010.
2. World Health Organization: Health and millennium development goals Geneva: World Health Organization; 2005.
3. World Health Organization, UNFPA and the World Bank: Trends in maternal mortality: 1990 to 2008 estimates developed by WHO, UNICEF, UNFPA and the World Bank Geneva: World Health Organization; 2010.
4. AbouZahr C. Global burden of maternal death and disability. Br Med Bull 2003.
5. Ethiopian Society of population studies, Maternal Health Care seeking Behaviour in Ethiopia using EDHS 2005, Addis Ababa, 2008.
6. Mekonnen, Yared, and Asnakech Mekonnen. Utilization of Maternal Health Care Services in Ethiopia. Calverton, Maryland, USA: ORC Macro; 2002.
7. World Health Organization (WHO). Improved access to maternal health services. WHO 98.7. Geneva: WHO; 1998.
8. World Bank. Better health in Africa: Experience and lessons learned. Washington, D.C.: World Bank; 1994a.
9. Fauveau, V., M.A. Koenig, J. Chakraborty, and A.I. Chowdhury. Causes of maternal mortality in rural Bangladesh: 1976-1985. Bulletin of the World Health Organization; 1988.
10. Fortney, J.A., I. Susanti, S. Gadalla, S. Saleh, P.J. Feldblum, and M. Potts. *Maternal mortality in Indonesian and Egypt*. British Journal of Gynaecology and Obstetrics; 1988.
11. Bhatia, S. Patterns and cause of neonatal and postneonatal mortality in rural Bangladesh. Studies in Family Planning; 1989.
12. CSA. Ethiopia Demographic and Health Survey 2005.Central statistics Agency, Addis Ababa. 2006.
13. CSA. Ethiopia Demographic and Health Survey 2000.Central Statistics Agency, Addis Ababa. 2001.
14. Central Statistical Agency (CSA) [Ethiopia] and ORC Macro. Ethiopia Demographic and Health Survey 2005. Addis Ababa, Ethiopia and Calverton, Maryland, USA: CSA and ORC Macro; 2006.
15. Anson O. Utilization of maternal care in rural HeiBei province, the People's Republic of China: individual & structural characteristics. Health Policy; 2004.

16. Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press; 1996.
17. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*; 1996.
18. Fayyad, U. & Uthurasamy, R. Data Mining and Knowledge Discovery in Databases" *Communications of the ACM*; 1996.
19. Lloyd-Williams, M., Williams, S, Bath, P. & Morris, C. Knowledge Discovery in the WHO *Health for All* Database: Developing A Taxonomy of Mortality Patterns for European Countries", In: Richards, B. & de Glanville,H. (eds) Current Perspectives in Healthcare Computing. Proceedings of HC96, pp. 551-556. Weybridge: BJHC; 1996.
20. Limb, P.R., & Meggs, G.J. Data Mining -Tools and Techniques. British Telecom Technology Journal; 1995.
21. Scarfe, R. & Shortland, R.J. Data Mining Applications in BT" IEE Colloquium on Knowledge Discovery in Databases. IEE Digest No: 1995/021(B); 1995.
22. Hedberg, S.R. The Data Gold Rush. *Byte*, 20(10), 83-88; 1995.
23. Watterson, K. "A Data Miner's Tools. *Byte* 20(10), 91-96. 4; 1995.
24. Lloyd-Williams, Michael. Discovering the hidden secrets in your data - the data mining approach to information. Information Research, **3**(2) Available at: http://informationr.net/ir/3-2/paper36.html;1997.
25. Ian H. Witten and Eibe Frank. Data mining Text book: Practical Machine Learning tools and Techniques. 2nd Ed. Morgan Kaufmann Publishers, San Francisco; 2005.
26. SPSS Inc. Neural Connection Applications Guide. Chicago: SPSS Inc; USA; 1995.
27. Jiawie, Han and Micheline Kamber. Data mining Concept and Techniques. 2nd Ed. Morgan Kaufmann Publishers, San Francisco; 2006.
28. S. P. Deshpande and V. M. Thakare. Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS) Volume 1, Number 1; 2010.
29. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. 3rd Ed. Two Crows Corporation. 500 Falls Road, Potomac, USA; 2005.
30. Mohammed Kantardzic J.B. Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc; 2003.

31. Han J and Kamber M. Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann; 2001.
32. Cios Krzysztof J., Pedrycz Wiltod., Swiniarski Roman W. Kurgan Lukasz A. Data Mining: A knowledge Discovery approach. New York: Springer-Verlag Science Business Media LLC; 2007.
33. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) CRISP-DM 1.0 Step-by-step data mining guide.
34. Obenshain, Mary K. *Application of data mining techniques to healthcare data: Infection control and hospital epidemiology.* The official journal of the Society of Hospital Epidemiologists of America 25 (8): 690-5; 2004.
35. Santos, Manuel Filipe,& Azevedo,Ana. KDD, SEMMA AND CRISP-DM: a parallel overview. IADIS European Confrence data mining; 2008.
36. Hian Cbye Kob and Gerald Tan. Data Mining Applications in Healthcare. Journal of Healthcare Information Management — Vol. 19, No. 2.
37. Van Lerberghe, W., and V. De Brouwere. Of blind alleys and things that have worked: history's lessons on reducing maternal mortality. In: De Brouwere, V. and W. Van Lerberghe, eds. Safe motherhood strategies: A recent review of the evidence. Antwerp, ITG Press, 7–33; 2001.
38. World Health Organization (WHO). Standards for maternal and neonatal care. Geneva: WHO; 2006.
39. WHO/UNICEF. Antenatal care in developing countries: promises, achievements and missed opportunities: an analysis of trends, levels and differentials, 1990-2001. 2003.
40. Bullough C, Meda N, Makowiecka K, Ronsmans C, Achadi EL, Hussein J. Current strategies for the reduction of maternal mortality. BJObG; 112(9):1180-8; 2005.
41. Berg CJ. Prenatal care in developing counties: the World Health Organization technical working group on antenatal care. J Am Med Womens Assoc 50(5):182-6; 1995.
42. De Bernis L, Sherratt DR, AbouZahr C, Van Lerberghe W. *Skilled attendants for pregnancy, childbirth and postnatal care*. Br Med Bull; 67:39-57; 2003.
43. Bloom SS, Lippeveld T, Wypij D. Does antenatal care make a difference to safe delivery? A study in urban Uttar Pradesh, India. Health Policy Plan; 14(1):38-48; 1999.

44. UNICEF. Millennium Development Goals, Goal: Improve maternal health. UNICEF. Available from: http://www.unicef.org/mdg/maternal.html; 2010.
45. Gill K, Pande R, Malhotra A. Women deliver for development. Lancet 13; 370 (9595):1347-57; 2007.
46. WHO. Maternal mortality ratio (per 100,000 live births). Available from: http://who.int/healthinfo/statistics/indmaternalmortality/en/index.html; 2010.
47. WHO. Making pregnancy safer. World Health Organization, Geneva. Available from: http://www.who.int/healthinfo/en/index.html; 2010.
48. Carlough M, McCall M. *Skilled birth attendance: what does it mean and how can it be measured? A clinical skills assessment of maternal and child health workers in Nepal*. Int J Gynaecol Obstet; 89(2):200-8; 2005.
49. Gubhaju MMaB. *Women's status, household structure and the Utilization of maternal health services in Nepal.* Asia-Pacific Population Journal 2001.
50. Celik Y, Hotchkiss DR. The socio-economic determinants of maternal health care utilization in Turkey. Soc Sci Med; 50(12):1797-806; 2000.
51. Woldemicael G, Tenkorang EY. *Women's Autonomy and Maternal Health Seeking Behavior in Ethiopia*. Matern Child Health J; 2009.
52. Mekonnen Y, Mekonnen A. *Factors influencing the use of maternal healthcare services in Ethiopia.* J Health Popul Nutr; 21(4):374-82; 2003.
53. Mesfin Nigussie DHM, Getnet Mitike. *Assessment of safe delivery service utilization among women of childbearing age in north Gondar Zone, North West Ethiopia*. Ethiop J Health Dev; 18(3); 2004.
54. Babalola S, Fatusi A. Determinants of use of maternal health services in Nigeria--looking beyond individual and household factors. BMC Pregnancy Childbirth; 9:43; 2009.
55. Z.Wang, S. Chen, and T. Sun. MultiK-MHKS: A novel multiple kernel learning algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30(3.2), 348–353, 2008.
56. Biset Desalegn. Predicting Low Birth Weight using Data Mining Techniques on Ethiopian Demography and Health Survey Dataset. School of Information Science. AAU; 2011.
57. Andersen, R. M. and Newman, J. F. Social and individual determinants of medical care utilization in the United States. Milbank Memorial Quarterly, 51, 95–124; 1973.

58. Han J and Kamber M. Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann; 2001.
59. Mohamed Kantardzic J.B. Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc; 2003.
60. Berry Michael W, Browne K M. Lecture Notes in Data Mining. USA: World Scientific Publishing Co. Pte. Ltd Inc. Rosewood Drive, Danvers, MA; 2006.
61. Ruben D and Canlas Jr. Data Mining in Healthcare: Current Application and Issues. Thesis, Australia: Carnegie Mellon University; 2009.
62. Berry Michael W, Browne K M. Lecture Notes in Data Mining. USA: World Scientific Publishing Co. Pte. Ltd Inc. Rosewood Drive, Danvers, MA; 2006.
63. Mohamed Kantardzic J.B. Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc; 2003.
64. Bouckaert R, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, Scuse D. Weka Manual for Version 3-6-0. New Zealand: University of Waikato, Hamilton; 2008.
65. Weiss Sholom M., Zhang Tong. Performance Analysis and Evaluation. In:Ye Nong, Editor. The Hand Book of Data Mining. New Jeresy. USA: Lawerence Erlbaum Associates Inc; 2003.
66. Ifeachor C E, Hamadicharef B. Receiver Operating Curve Analysis in The Evaluation of Intelligent Medical Systems. UK: University of Plymouth. Drake Circus Plymouth PL4 8AA, Devon; 2004.
67. Melanie C. Page, Sanford L Braver, David P. MacKinnon. Levine's Guide to SPSS for Analysis of Variance. London: Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey; 2003.
68. Cios Krzysztof J., Pedrycz Wiltod., Swiniarski Roman W. Kurgan Lukasz A. Data Mining: A knowledge Discovery approach. New York: Springer-Verlag Science Business Media LLC; 2007.
69. Daniel T. Larose. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, New Jersey; 2005.
70. Ethiopia Demographic and Health Survey 2011 report Central Statistical Agency ,Addis Ababa, Ethiopia; 2011.
71. Colin Shearer. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Ware Housing Volume 5, Number 4; 2000.

Appendices

Appendix A: J48 Decision tree classifier output for ANC

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.18 -M 5

Relation: ANC-weka.filters.unsupervised.attribute.ReplaceMissingValues-unset-class-temporarily

Instances: 7764

Attributes: 13

Age of women's

Region

Place of Residence

Highest educational level

Religion

Media Exposure

Wealth Index

Total children born

Current marital status

Husband's education level

Husband's occupation

Respondent's occupation

Antenatal care

Test mode:10-fold cross-validation

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

J48 pruned tree

Place of Residence = Rural

| Highest educational level = Primary

| | Husbands occupation = Agricultural - employee: No (1173.0/436.0)

| | Husbands occupation = Non-agricultural - employee

| | | Religion = Orthodox: Yes (82.0/25.0)

| | | Religion = Muslim

| | | | Current marital status = never married: Yes (0.0)

| | | | Current marital status = Living with partner

| | | | | Wealth Index = Poorest: No (8.0/4.0)

| | | | | Wealth Index = Middle: No (14.0/4.0)

| | | | | Wealth Index = Poorer: No (9.0/4.0)

| | | | | Wealth Index = Richest: Yes (24.0/7.0)
| | | | | Wealth Index = Richer: No (23.0/8.0)
| | | | Current marital status = Not living together: Yes (8.0/1.0)
| | | Religion = Protestant
| | | | Region = Amhara: No (0.0)
| | | | Region = Addis Ababa: No (0.0)
| | | | Region = Harari: No (0.0)
| | | | Region = Somali: No (0.0)
| | | | Region = Benishangul-Gumuz: No (1.0)
| | | | Region = Oromiya: Yes (11.0/3.0)
| | | | Region = SNNP: No (41.0/12.0)
| | | | Region = Tigray: No (0.0)
| | | | Region = Affar: No (0.0)
| | | | Region = Gambela: Yes (49.0/23.0)
| | | | Region = Dire Dawa: No (0.0)
| | | Religion = Other: Yes (1.0)
| | | Religion = Catholic: No (8.0/1.0)
| | | Religion = Traditional: Yes (1.0)
| | Husbands occupation = Did not work: No (44.0/14.0)
| Highest educational level = No education: No (4665.0/1065.0)
| Highest educational level = Secondary
| | Husbands occupation = Agricultural - employee: No (26.0/12.0)
| | Husbands occupation = Non-agricultural - employee: Yes (30.0/4.0)
| | Husbands occupation = Did not work: No (3.0/1.0)
| Highest educational level = Higher: Yes (30.0/5.0)
Place of Residence = Urban
| Wealth Index = Poorest
| | Highest educational level = Primary: Yes (20.0/6.0)
| | Highest educational level = No education: No (38.0/13.0)
| | Highest educational level = Secondary: Yes (3.0)
| | Highest educational level = Higher: Yes (1.0)
| Wealth Index = Middle: No (17.0/7.0)
| Wealth Index = Poorer: No (13.0/3.0)
| Wealth Index = Richest: Yes (1320.0/229.0)
| Wealth Index = Richer
| | Husbands education level = No education
| | | Religion = Orthodox: Yes (10.0/3.0)

| | | Religion = Muslim: No (25.0/6.0)
| | | Religion = Protestant: No (5.0/1.0)
| | | Religion = Other: No (0.0)
| | | Religion = Catholic: No (0.0)
| | | Religion = Traditional: No (0.0)
| | Husbands education level = Primary: Yes (46.0/18.0)
| | Husbands education level = Secondary: Yes (10.0/3.0)
| | Husbands education level = Higher: Yes (5.0)

Number of Leaves: 45
Size of the tree: 57

Time taken to build model: 0.05 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	5806	74.781 %
Incorrectly Classified Instances	1958	25.219 %
Kappa statistic	0.4106	
Mean absolute error	0.3661	
Root mean squared error	0.4303	
Relative absolute error	78.0575 %	
Root relative squared error	88.8672 %	
Total Number of Instances	7764	

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.93	0.556	0.736	0.93	0.822	0.717	No
	0.444	0.07	0.793	0.444	0.569	0.717	Yes
Weighted Avg.	0.748	0.373	0.757	0.748	0.727	0.717	

=== Confusion Matrix ===

a	b	<-- classified as
4511	338	a = No
1620	1295	b = Yes

Appendix B: J48 Decision tree classifier output for Delivery care

=== Run information ===

Scheme: weka.classifiers.trees.J48 -U -M 10

Relation: Assistance-weka.filters.unsupervised.attribute.ReplaceMissingValues-unset-class-temporarily-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1-weka.filters.unsupervised.attribute.Remove-R9-10,12

Instances: 7764

Attributes: 10

Place of Residence

Wealth Index

Husbands/partner's occupation

Highest educational level

Region

Husbands/Partner's education level

Media Exposure

Total children born

Age of women's

Assistance at delivery

Test mode:10-fold cross-validation

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

J48 unpruned tree

Place of Residence = Rural

| Highest educational level = Primary

| | Husbands/partner's occupation = Agricultural - employee

| | | Total children born = Two or Three Child

| | | | Husbands/Partner's education level = No education: No (136.0/11.0)

| | | | Husbands/Partner's education level = Primary: No (215.0/13.0)

| | | | Husbands/Partner's education level = Secondary

| | | | | Wealth Index = Poorest: No (10.0/1.0)

| | | | | Wealth Index = Middle: No (1.0)

| | | | | Wealth Index = Poorer: No (3.0)

| | | | | Wealth Index = Richest: Yes (1.0)

| | | | | Wealth Index = Richer: No (10.0/3.0)

| | | | Husbands/Partner's education level = Higher: No (5.0/1.0)

| | | Total children born = Six or more child: No (245.0/6.0)

| | | Total children born = Four or Five Child: No (230.0/10.0)

| | | Total children born = One child

| | | | Region = Amhara: No (41.0/4.0)

| | | | Region = Addis Ababa: No (0.0)

| | | | Region = Harari: No (13.0/5.0)

| | | | Region = Somali: No (2.0)

| | | | Region = Benishangul-Gumuz: No (44.0/4.0)

| | | | Region = Oromiya

| | | | | Husbands/Partner's education level = No education: No (17.0/2.0)

| | | | | Husbands/Partner's education level = Primary: No (39.0/5.0)

| | | | | Husbands/Partner's education level = Secondary: No (3.0)

| | | | | Husbands/Partner's education level = Higher: Yes (1.0)

| | | | Region = SNNP: No (50.0/2.0)

| | | | Region = Tigray: No (57.0/4.0)

| | | | Region = Affar: No (3.0/1.0)

| | | | Region = Gambela

| | | | | Wealth Index = Poorest: No (16.0/1.0)

| | | | | Wealth Index = Middle: No (8.0/1.0)

| | | | | Wealth Index = Poorer: No (3.0)

| | | | | Wealth Index = Richest: Yes (1.0)

| | | | | Wealth Index = Richer: No (13.0/5.0)

| | | | Region = Dire Dawa: No (6.0/1.0)

| | Husbands/partner's occupation = Non-agricultural - employee

| | | Media Exposure = No media access

| | | | Region = Amhara: No (7.0/2.0)

| | | | Region = Addis Ababa: No (0.0)

| | | | Region = Harari: No (2.0)

| | | | Region = Somali: No (11.0/1.0)

| | | | Region = Benishangul-Gumuz: No (14.0/1.0)

| | | | Region = Oromiya: No (26.0/1.0)

| | | | Region = SNNP: No (37.0/1.0)

| | | | Region = Tigray: No (22.0/2.0)

| | | | Region = Affar: No (4.0)

| | | | Region = Gambela: No (68.0/9.0)

| | | | Region = Dire Dawa: Yes (3.0/1.0)

| | | Media Exposure = Media Access

| | | | Region = Amhara: No (5.0/2.0)

| | | | Region = Addis Ababa: No (0.0)

| | | | Region = Harari: Yes (2.0)

| | | | Region = Somali: No (4.0/2.0)

| | | | Region = Benishangul-Gumuz: No (9.0/3.0)

| | | | Region = Oromiya: No (17.0/3.0)

| | | | Region = SNNP: No (21.0/3.0)

| | | | Region = Tigray: No (17.0/6.0)

| | | | Region = Affar: No (1.0)

| | | | Region = Gambela: Yes (6.0/2.0)

| | | | Region = Dire Dawa: Yes (4.0/1.0)

| | Husbands/partner's occupation = Did not work: No (44.0/5.0)

| Highest educational level = No education

| | Husbands/partner's occupation = Agricultural - employee

| | | Wealth Index = Poorest: No (1634.0/22.0)

| | | Wealth Index = Middle

| | | | Total children born = Two or Three Child: No (251.0/8.0)

| | | | Total children born = Six or more child: No (266.0/5.0)

| | | | Total children born = Four or Five Child: No (213.0/6.0)

| | | | Total children born = One child

| | | | | Region = Amhara: No (25.0/2.0)

| | | | | Region = Addis Ababa: No (0.0)

| | | | | Region = Harari: No (7.0/1.0)

| | | | | Region = Somali: No (2.0)

| | | | | Region = Benishangul-Gumuz: No (5.0)

| | | | | Region = Oromiya: No (18.0/1.0)

| | | | | Region = SNNP: No (11.0)

| | | | | Region = Tigray: No (6.0)

| | | | | Region = Affar: No (1.0)

| | | | | Region = Gambela: Yes (1.0)

| | | | | Region = Dire Dawa: No (6.0/1.0)

| | | Wealth Index = Poorer: No (950.0/23.0)

| | | Wealth Index = Richest: No (107.0/10.0)

| | | Wealth Index = Richer: No (587.0/28.0)

| | Husbands/partner's occupation = Non-agricultural - employee: No (513.0/30.0)

| | Husbands/partner's occupation = Did not work: No (62.0/4.0)

| Highest educational level = Secondary: No (59.0/21.0)

| Highest educational level = Higher: Yes (30.0/13.0)

Place of Residence = Urban

| Highest educational level = Primary

| | Wealth Index = Poorest: No (20.0/7.0)

| | Wealth Index = Middle: No (7.0/1.0)

| | Wealth Index = Poorer: No (2.0/1.0)

| | Wealth Index = Richest

| | | Region = Amhara: Yes (25.0/8.0)

| | | Region = Addis Ababa

| | | | Husbands/partner's occupation = Agricultural - employee: No (11.0/5.0)

| | | | Husbands/partner's occupation = Non-agricultural - employee

| | | | | Age of women's = 20 - 34: Yes (118.0/17.0)

| | | | | Age of women's = 35+: Yes (21.0/3.0)

| | | | | Age of women's = <= 19: No (2.0)

| | | | Husbands/partner's occupation = Did not work: Yes (0.0)

| | | Region = Harari

| | | | Total children born = Two or Three Child: Yes (31.0/6.0)

| | | | Total children born = Six or more child: No (2.0)

| | | | Total children born = Four or Five Child: No (7.0/3.0)

| | | | Total children born = One child: Yes (24.0/4.0)

| | | Region = Somali

| | | | Media Exposure = No media access: No (10.0/2.0)

| | | | Media Exposure = Media Access: Yes (15.0/5.0)

| | | Region = Benishangul-Gumuz: Yes (12.0/4.0)

| | | Region = Oromiya

| | | | Total children born = Two or Three Child

| | | | | Media Exposure = No media access: No (10.0/4.0)

| | | | | Media Exposure = Media Access: Yes (14.0/4.0)

| | | | Total children born = Six or more child: No (6.0)

| | | | Total children born = Four or Five Child: No (6.0/3.0)

| | | | Total children born = One child: Yes (10.0/2.0)

| | | Region = SNNP

| | | | Total children born = Two or Three Child: No (5.0/2.0)

| | | | Total children born = Six or more child: Yes (4.0/1.0)

| | | | Total children born = Four or Five Child: No (15.0/4.0)

| | | | Total children born = One child: Yes (14.0/4.0)

| | | Region = Tigray

| | | | Husbands/Partner's education level = No education: No (9.0)

| | | | Husbands/Partner's education level = Primary: Yes (19.0/6.0)

| | | | Husbands/Partner's education level = Secondary: Yes (15.0/5.0)

| | | | Husbands/Partner's education level = Higher: Yes (6.0/1.0)

| | | Region = Affar

| | | | Media Exposure = No media access: No (10.0/2.0)

| | | | Media Exposure = Media Access: Yes (27.0/11.0)

| | | Region = Gambela: Yes (19.0/3.0)

| | | Region = Dire Dawa: Yes (71.0/7.0)

| | Wealth Index = Richer: No (31.0/12.0)

| Highest educational level = No education

| | Wealth Index = Poorest: No (38.0/4.0)

| | Wealth Index = Middle: No (9.0)

| | Wealth Index = Poorer: No (10.0/2.0)

| | Wealth Index = Richest

| | | Region = Amhara: No (33.0/6.0)

| | | Region = Addis Ababa

| | | | Total children born = Two or Three Child: Yes (27.0/6.0)

| | | | Total children born = Six or more child: No (5.0/1.0)

| | | | Total children born = Four or Five Child: Yes (15.0/5.0)

| | | | Total children born = One child

| | | | | Media Exposure = No media access: No (11.0/4.0)

| | | | | Media Exposure = Media Access: Yes (14.0/3.0)

| | | Region = Harari

| | | | Total children born = Two or Three Child: Yes (18.0/6.0)

| | | | Total children born = Six or more child: No (2.0)

| | | | Total children born = Four or Five Child: Yes (10.0/2.0)

| | | | Total children born = One child: Yes (8.0)

| | | Region = Somali

| | | | Age of women's = 20 - 34

| | | | | Total children born = Two or Three Child: No (14.0/4.0)

| | | | | Total children born = Six or more child: No (17.0/3.0)

| | | | | Total children born = Four or Five Child: No (7.0/1.0)

| | | | | Total children born = One child: Yes (4.0/1.0)

| | | | Age of women's = 35+: No (13.0)

| | | | Age of women's = <= 19: No (2.0)

| | | Region = Benishangul-Gumuz: No (11.0/4.0)

| | | Region = Oromiya: No (23.0/9.0)

| | | Region = SNNP: No (5.0/1.0)

| | | Region = Tigray: No (30.0/9.0)

| | | Region = Affar: No (30.0/5.0)

| | | Region = Gambela: No (13.0/6.0)

| | | Region = Dire Dawa: Yes (84.0/20.0)

| | Wealth Index = Richer

| | | Husbands/Partner's education level = No education: No (34.0/4.0)

| | | Husbands/Partner's education level = Primary: No (28.0/8.0)

| | | Husbands/Partner's education level = Secondary: No (3.0)

| | | Husbands/Partner's education level = Higher: Yes (1.0)

| Highest educational level = Secondary: Yes (253.0/35.0)

| Highest educational level = Higher: Yes (143.0/12.0)

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.961	0.335	0.936	0.961	0.948	0.884	No
	0.665	0.039	0.771	0.665	0.714	0.884	Yes
Weighted Avg.	0.913	0.286	0.909	0.913	0.91	0.884	

=== Confusion Matrix ===

a	b	<-- classified as
6239	252	a = No
426	847	b = Yes

Appendix C: J48 Decision tree classifier output for Postnatal care

=== Run information ===

Scheme: weka.classifiers.trees.J48 -U -M 10

Relation: Postnatal care-weka.filters.unsupervised.attribute.ReplaceMissingValues-unset-class-temporarily-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1-weka.filters.unsupervised.attribute.Remove-R9-12

Instances: 7764

Attributes: 9

Wealth Index

Place of Residence

Husbands/partner's occupation

Highest educational level

Husbands/Partner's education level

Region

Media Exposure

Total children born

Postnatal care

Test mode:10-fold cross-validation

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

J48 unpruned tree

Place of Residence = Rural

| Highest educational level = Primary

| | Husbands/partner's occupation = Agricultural - employee

| | | Region = Amhara: No (127.0/4.0)

| | | Region = Addis Ababa: No (0.0)

| | | Region = Harari

| | | | Media Exposure = No media access

| | | | | Husbands/Partner's education level = No education: No (17.0/2.0)

| | | | | Husbands/Partner's education level = Primary: No (24.0/3.0)

| | | | | Husbands/Partner's education level = Secondary: No (1.0)

| | | | | Husbands/Partner's education level = Higher: Yes (1.0)

| | | | Media Exposure = Media Access: No (16.0/7.0)

| | | Region = Somali: No (21.0)

| | | Region = Benishangul-Gumuz

| | | | Husbands/Partner's education level = No education

| | | | | Wealth Index = Poorest: No (8.0/1.0)

| | | | | Wealth Index = Middle: No (16.0/1.0)

| | | | | Wealth Index = Poorer: No (10.0)

| | | | | Wealth Index = Richest: Yes (1.0)

| | | | | Wealth Index = Richer: No (9.0/3.0)

| | | | Husbands/Partner's education level = Primary: No (68.0/5.0)

| | | | Husbands/Partner's education level = Secondary: No (7.0)

| | | | Husbands/Partner's education level = Higher: No (1.0)

| | | Region = Oromiya: No (253.0/14.0)

| | | Region = SNNP: No (259.0/10.0)

| | | Region = Tigray

| | | | Wealth Index = Poorest: No (53.0/8.0)

| | | | Wealth Index = Middle: No (30.0/2.0)

| | | | Wealth Index = Poorer: No (36.0/3.0)

| | | | Wealth Index = Richest: No (4.0/1.0)

| | | | Wealth Index = Richer

| | | | | Husbands/Partner's education level = No education: No (12.0/3.0)

| | | | | Husbands/Partner's education level = Primary: No (17.0/2.0)

| | | | | Husbands/Partner's education level = Secondary: Yes (1.0)

| | | | | Husbands/Partner's education level = Higher: No (0.0)

| | | Region = Affar: No (10.0/1.0)

| | | Region = Gambela

| | | | Wealth Index = Poorest: No (74.0/6.0)

| | | | Wealth Index = Middle: No (19.0/2.0)

| | | | Wealth Index = Poorer: No (17.0/1.0)

| | | | Wealth Index = Richest: Yes (5.0/2.0)

| | | | Wealth Index = Richer: No (34.0/6.0)

| | | Region = Dire Dawa: No (22.0/2.0)

| | Husbands/partner's occupation = Non- agricultural - employee

| | | Region = Amhara: No (12.0/3.0)

| | | Region = Addis Ababa: No (0.0)

| | | Region = Harari: No (4.0/2.0)

| | | Region = Somali: No (15.0/2.0)

| | | Region = Benishangul-Gumuz: No (23.0/4.0)

| | | Region = Oromiya: No (43.0/4.0)

| | | Region = SNNP: No (58.0/6.0)

| | | Region = Tigray

| | | | Wealth Index = Poorest: No (3.0/1.0)

| | | | Wealth Index = Middle: Yes (6.0/2.0)

| | | | Wealth Index = Poorer: No (6.0)

| | | | Wealth Index = Richest: No (14.0/5.0)

| | | | Wealth Index = Richer: No (10.0/1.0)

| | | Region = Affar: No (5.0)

| | | Region = Gambela: No (74.0/14.0)

| | | Region = Dire Dawa: No (7.0/1.0)

| | Husbands/partner's occupation = Did not work: No (44.0/3.0)

| Highest educational level = No education

| | Media Exposure = No media access

| | | Husbands/Partner's education level = No education: No (2897.0/106.0)

| | | Husbands/Partner's education level = Primary: No (1123.0/60.0)

| | | Husbands/Partner's education level = Secondary

| | | | Region = Amhara: No (2.0)

| | | | Region = Addis Ababa: No (0.0)

| | | | Region = Harari: No (2.0)

| | | | Region = Somali: No (8.0)

| | | | Region = Benishangul-Gumuz: No (3.0)

| | | | Region = Oromiya: No (8.0/1.0)

| | | | Region = SNNP: No (13.0/1.0)

| | | | Region = Tigray: Yes (3.0/1.0)

| | | | Region = Affar: No (7.0)

| | | | Region = Gambela: No (18.0/2.0)

| | | | Region = Dire Dawa: No (0.0)

| | | Husbands/Partner's education level = Higher: No (17.0)

| | Media Exposure = Media Access

| | | Husbands/partner's occupation = Agricultural - employee

| | | | Region = Amhara: No (83.0/5.0)

| | | | Region = Addis Ababa: No (0.0)

| | | | Region = Harari

| | | | | Wealth Index = Poorest: No (0.0)

| | | | | Wealth Index = Middle: Yes (2.0)

| | | | | Wealth Index = Poorer: No (0.0)

| | | | | Wealth Index = Richest: No (14.0/3.0)

| | | | | Wealth Index = Richer: No (11.0/1.0)

| | | | Region = Somali: No (21.0)

| | | | Region = Benishangul-Gumuz: No (41.0/2.0)

| | | | Region = Oromiya: No (110.0/4.0)

| | | | Region = SNNP: No (72.0/5.0)

| | | | Region = Tigray: No (53.0/7.0)

| | | | Region = Affar: No (41.0/1.0)

| | | | Region = Gambela: No (4.0)

| | | | Region = Dire Dawa: No (8.0/1.0)

| | | Husbands/partner's occupation = Non-agricultural - employee: No (95.0/17.0)

| | | Husbands/partner's occupation = Did not work: No (9.0)

| Highest educational level = Secondary

| | Total children born = Two or Three Child: No (24.0/4.0)

| | Total children born = Six or more child: Yes (1.0)

| | Total children born = Four or Five Child: No (2.0/1.0)

| | Total children born = One child

| | | Husbands/Partner's education level = No education: No (2.0)

| | | Husbands/Partner's education level = Primary: No (10.0/2.0)

| | | Husbands/Partner's education level = Secondary: No (9.0/3.0)

| | | Husbands/Partner's education level = Higher: Yes (11.0/5.0)

| Highest educational level = Higher: Yes (30.0/11.0)

Place of Residence = Urban

| Highest educational level = Primary

| | Wealth Index = Poorest: No (20.0/7.0)

| | Wealth Index = Middle: No (7.0)

| | Wealth Index = Poorer: No (2.0/1.0)

| | Wealth Index = Richest

| | | Husbands/Partner's education level = No education

| | | | Region = Amhara: No (2.0)

| | | | Region = Addis Ababa: No (26.0/7.0)

| | | | Region = Harari: No (16.0/7.0)

| | | | Region = Somali: No (6.0)

| | | | Region = Benishangul-Gumuz: No (4.0/1.0)

| | | | Region = Oromiya: No (7.0/2.0)

| | | | Region = SNNP: No (5.0/1.0)

| | | | Region = Tigray: No (9.0/1.0)

| | | | Region = Affar: No (9.0/3.0)

| | | | Region = Gambela: Yes (2.0)

| | | | Region = Dire Dawa: No (4.0/1.0)

| | | Husbands/Partner's education level = Primary

| | | | Region = Amhara: No (15.0/7.0)

| | | | Region = Addis Ababa

| | | | | Total children born = Two or Three Child: No (30.0/13.0)

| | | | | Total children born = Six or more child: Yes (5.0/1.0)

| | | | | Total children born = Four or Five Child: Yes (7.0)

| | | | | Total children born = One child: No (40.0/12.0)

| | | | Region = Harari

| | | | | Total children born = Two or Three Child: Yes (15.0/7.0)

| | | | | Total children born = Six or more child: No (2.0)

| | | | | Total children born = Four or Five Child: No (4.0/1.0)

| | | | | Total children born = One child: Yes (10.0)

| | | | Region = Somali: No (7.0/2.0)

| | | | Region = Benishangul-Gumuz: Yes (6.0)

| | | | Region = Oromiya: No (21.0/2.0)

| | | | Region = SNNP: No (22.0/9.0)

| | | | Region = Tigray: No (19.0/8.0)

| | | | Region = Affar: No (14.0/6.0)

| | | | Region = Gambela: Yes (11.0/5.0)

| | | | Region = Dire Dawa: No (29.0/9.0)

| | | Husbands/Partner's education level = Secondary

| | | | Region = Amhara: No (6.0/3.0)

| | | | Region = Addis Ababa

| | | | | Total children born = Two or Three Child: No (12.0/5.0)

| | | | | Total children born = Six or more child: Yes (1.0)

| | | | | Total children born = Four or Five Child: Yes (5.0/2.0)

| | | | | Total children born = One child: Yes (19.0/8.0)

| | | | Region = Harari: Yes (10.0/3.0)

| | | | Region = Somali: Yes (10.0/4.0)

| | | | Region = Benishangul-Gumuz: Yes (2.0)

| | | | Region = Oromiya: No (11.0/4.0)

| | | | Region = SNNP: No (5.0/2.0)

| | | | Region = Tigray: Yes (15.0/5.0)

| | | | Region = Affar: No (8.0/3.0)

| | | | Region = Gambela: No (4.0/2.0)

| | | | Region = Dire Dawa: Yes (29.0/10.0)

| | | Husbands/Partner's education level = Higher: Yes (54.0/24.0)

| | Wealth Index = Richer: No (31.0/5.0)

| Highest educational level = No education

| | Region = Amhara: No (42.0/5.0)

| | Region = Addis Ababa

| | | Total children born = Two or Three Child

| | | | Media Exposure = No media access: Yes (11.0/4.0)

| | | | Media Exposure = Media Access: No (16.0/7.0)

| | | Total children born = Six or more child: No (5.0/1.0)

| | | Total children born = Four or Five Child: No (15.0/6.0)

| | | Total children born = One child: No (26.0/9.0)

| | Region = Harari

| | | Husbands/Partner's education level = No education: Yes (13.0/6.0)

| | | Husbands/Partner's education level = Primary: No (21.0/8.0)

| | | Husbands/Partner's education level = Secondary: Yes (8.0/3.0)

| | | Husbands/Partner's education level = Higher: No (2.0/1.0)

| | Region = Somali: No (99.0/14.0)

| | Region = Benishangul-Gumuz: No (33.0/5.0)

| | Region = Oromiya

| | | Husbands/Partner's education level = No education: No (10.0/1.0)

| | | Husbands/Partner's education level = Primary: No (18.0/4.0)

| | | Husbands/Partner's education level = Secondary: Yes (3.0/1.0)

| | | Husbands/Partner's education level = Higher: Yes (1.0)

| | Region = SNNP: No (8.0)

| | Region = Tigray: No (33.0/13.0)

| | Region = Affar: No (41.0/7.0)

| | Region = Gambela: Yes (22.0/8.0)

| | Region = Dire Dawa: No (92.0/29.0)

| Highest educational level = Secondary

| | Region = Amhara: Yes (6.0/2.0)

| | Region = Addis Ababa: Yes (80.0/23.0)

| | Region = Harari: Yes (40.0/11.0)

| | Region = Somali: No (10.0/5.0)

| | Region = Benishangul-Gumuz: Yes (7.0/1.0)

| | Region = Oromiya: No (17.0/8.0)

| | Region = SNNP: Yes (12.0/5.0)

| | Region = Tigray: Yes (25.0/9.0)

| | Region = Affar: Yes (14.0/6.0)

| | Region = Gambela: No (10.0/3.0)

| | Region = Dire Dawa: Yes (32.0/11.0)

| Highest educational level = Higher

| | Region = Amhara: Yes (9.0/4.0)

| | Region = Addis Ababa

| | | Total children born = Two or Three Child: Yes (15.0/5.0)

| | | Total children born = Six or more child: Yes (0.0)

| | | Total children born = Four or Five Child: Yes (2.0)

| | | Total children born = One child

| | | | Husbands/Partner's education level = No education: Yes (0.0)

| | | | Husbands/Partner's education level = Primary: No (3.0/1.0)

| | | | Husbands/Partner's education level = Secondary: Yes (10.0/2.0)

| | | | Husbands/Partner's education level = Higher: Yes (13.0/3.0)

| | Region = Harari: Yes (27.0/6.0)

| | Region = Somali: Yes (1.0)

| | Region = Benishangul-Gumuz: No (4.0/2.0)

| | Region = Oromiya: No (13.0/5.0)

| | Region = SNNP: Yes (7.0/3.0)

| | Region = Tigray: Yes (7.0)

| | Region = Affar: Yes (5.0/2.0)

| | Region = Gambela: Yes (10.0/2.0)

| | Region = Dire Dawa: Yes (17.0/4.0)

Number of Leaves: 184

Size of the tree: 223

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	6797	87.5451 %
Incorrectly Classified Instances	967	12.4549 %
Kappa statistic	0.3592	
Mean absolute error	0.1742	
Root mean squared error	0.306	
Relative absolute error	73.5895 %	
Root relative squared error	88.9318 %	
Total Number of Instances	7764	