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

GRADUATE PROGRAM IN HEALTH INFORMATICS

PREDICTING LOW BIRTH WEIGHT USING DATA MINING TECHNIQUES ON ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY DATA SETS

BY

BISET DESALEGN

JUNE 2011
ADDIS ABABA, ETHIOPIA

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

GRADUATE PROGRAM IN HEALTH INFORMATICS

PREDICTING LOW BIRTH WEIGHT USING DATA MINING TECHNIQUES ON ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY DATA SETS

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

BISET DESALEGN

JUNE 2011

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

JOINT PROGRAM BETWEEN SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

PREDICTING LOW BIRTH WEIGHT USING DATA MINING TECHNIQUES ON ETHIOPIA DEMOGRAPHIC AND HEALTH SURVEY DATA SETS

BY

BISET DESALEGN

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	_____
Dr. Dereje Teferi	Advisor,	_____	_____
Dr. Mitike Molla	Advisor,	_____	_____
Dr. Million Meshesha	Examiner,	_____	_____

Declaration

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

Name: _____

Signature: _____

Date: _____

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

Name:	_____	**Name:**	_____
Signature:	_____	**Signature:**	_____
Date:	_____	**Date:**	_____

ACKNOWLEDGMENTS

I would like to thank all those who have helped me to accomplish this thesis. First, I gratefully express my deepest thanks to the almighty God for his guidance, help, support and who added days to my age, Glory to God. Next, I would like to extend my deepest thanks to my advisors Dr. Dereje Teferi and Dr. Mitike Molla for their wonderful and unreserved assistance in every step of this study. Thirdly, I would like to thanks central statistics agency and Measure DHS for giving Low birth weight dataset.

I am thankful for the support of my colleagues Solomon G/meskel, Abebech Haile, Abel Damtaw, Anteneh Fetene, Geletaw Sahle, Haftom G/egiziabher and Selam Assamnew.

Finally, I am also thankful to my parents (Tenagne W/semayat and Eshetu Mihretu), brothers (Ayenew, Kidem, Mihret, Habtam, Biniam and Wube), sisters (Messi, Marta and Bizuhan) for their encouragement and support. Their all rounded and absolute support enabled me to realize my educational goal.

Tables of Contents

Contents **Page**

LISTS OF FIGURE

LISTS OF TABLES

LIST OF ACRONYMS

ARFF:	Attribute Relation File Format
BMI:	Body-Mass Index
CART:	Classification and Regression Tree
CHAID:	Chai-squared Automation Interaction Detection
CRISP-DM:	CRoss-Industry Standard Process for Data Mining
CSA:	Central Statistical Agency
CSV:	Comma Separated Value
DNBC:	Danish National Birth Cohort
EDHS:	Ethiopia Demographic and Health Survey
IBM:	International Business Machine
KDD:	Knowledge Discovery in Databases
LBW:	Low Birth Weight
MDG:	Millennium Development Goal
OLAP:	On-Line Analytical Processing
PHCCO:	Population and Housing Census Commission Office
PASDEP	Plan for Accelerated and Sustained Development to End Poverty
SPSS:	Statistical Package for the Social Sciences
UNICEF:	United Nations Children's Fund
WEKA:	Waikato Environment for Knowledge Analysis
WHO:	World Health Organization

ABSTRACT

Low birth weight is one of the critical issues in Ethiopia that causes many babies short-term and long-term health consequences and tend to have higher mortality and morbidity. DHS Ethiopia report shows that the percentage of low birth weight babies has increased from 8 percent in 2000 to 14 percent in 2005. The percentage of babies assessed by mothers as being very small at birth has increased over the same period from 6 percent to 21 percent

Low birth weight is a reasonable well-defined problem caused by many factors that are potentially modifiable and the costs of preventing them are well within reach, even in poor countries like Ethiopia. Therefore, it is very important to predict LBW in various communities in the country in order to come up with feasible intervention strategies to minimize the problem. Data mining techniques is a good tool to explore hidden knowledge from huge data set. The goal of this study is to predict low birth weight using EDHS 2005 (Ethiopia Demographic Health Survey) data set by applying of data mining technology. This study tried to build a model using data mining technique addressing the factors associated with low birth weight.

To this end, data was collected from Measure DHS Ethiopia. The methodology applied in this research was CRISP-DM, which contains six major phases: business understanding, data understanding, and data preparation, model building, evaluation and deployment. A total of 9861 records were used for the experiments. Some attributes of numeric values are discretized using ten-bin discretization implemented in Weka. Besides, missing values are also handling by data imputation technique, which replaces all missing values with the modes for nominal and categorical and means for numerical values from the training instances.

The selected data mining techniques for predicting low birth weight was classification. J48 decision tree classifier and PART rule induction algorithms were selected for experiments.

Several models were built implementing the J48 decision tree classifier algorithm and PART rule induction algorithms. These experiments has been done using pruning with all and reduced attributes, by giving J48 classifiers parameters in different values. The researcher compare the classification performance of the decision trees with tree prunning and without tree prunning, and found that tree prunning can significantly improve decision tree's classification performance.

In general, the results from this study were encouraging; it can be used as decision support aid for health practitioner. 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.

CHAPTER ONE

INTRODUCTION

1.1 Background

Low birth weight (LBW) is the main factor determining neonatal and prenatal survival and it is also associated with many adverse outcomes in newborns (Kramer, 1987). Low birth weight has been defined by the World Health Organization (WHO) as weight at birth of less than 2,500 grams (WHO Report, 1992). This practical cut-off for international comparison is based on epidemiological observations that infants weighing less than 2,500 grams are approximately 20 times more likely to die than heavier babies (Kramer, 1987). More common in developing than developed countries, a birth weight below 2,500 grams contributes to a range of poor health outcomes (UNICEF/WHO, 2004).

According to WHO report, reducing low birth weight incidence by at least one third between 2000 and 2010 is one of the major goals in 'A World Fit for Children,' the Declaration and Plan of Action adopted at the United Nations General Assembly Special Session on Children in 2002. The reduction of low birth weight also forms an important contribution to the Millennium Development Goal (MDG) for reducing child mortality. Activities towards the achievement of the MDGs will need to ensure a healthy start in life for children by making certain that women commence pregnancy healthy and well nourished, and go through pregnancy and childbirth safely. Low birth weight is therefore an important indicator for monitoring progress towards these internationally agreed-upon goals (UNICEF/WHO, 2004).

Low birth weight is one of the critical issues in Ethiopia that causes many babies short-term and long-term health consequences and tend to have higher mortality and morbidity. DHS Ethiopia /2005/ report shows that the percentage of low birth weight babies has increased in the past five years from 8 percent in 2000 to 14 percent in 2005. The percentage of babies assessed by mothers as being very small at birth has increased over

the same period from 6 percent to 21 percent (Ethiopia Demographic and Health Survey, 2005).

Low birth weight has long been used as an important public health indicator. Low birth weight is not a proxy for any one dimension of either maternal or prenatal health outcomes. Globally, the indicator is a good summary measure of a multifaceted public health problem that includes long-term maternal malnutrition, ill health, hard work and poor pregnancy health care. Countries should therefore be encouraged to ensure accurate and reliable weighing of infants as close to birth as possible (UNICEF/WHO, 2004).

Low birth weight (LBW) can be caused either by premature delivery (short gestation<37 week) or by foetal growth retardation. Known factors for pre-term delivery and foetal growth retardation which are associated with LBW include low maternal food intake, hard physical work during pregnancy, and illness, especially infections. The studies suggest that cigarette smoking, genetic and environmental factors can cause LBW, short maternal stature, very young age, high parity, close birth spacing is all associated factors (Kramer, 1987; UNICEF/WHO, 2004).

Low birth weight is a reasonable well-defined problem caused by factors that are potentially modifiable and the costs of preventing them are well within reach, even in poor countries like Ethiopia. Therefore, it is very important to predict LBW in various communities in the country in order to come up with feasible intervention strategies to minimize the problem.

According to Moheb, et al. (2005) researchers take many hours to collect information into books and matrixes to try to figure out the cause of disease. With data mining the medical healthcare industry was able to characterize patient behavior to predict office visits, Identify successful medical therapies for different illnesses, and Analyze cause and effect of diseases. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making (Koh and Tan, 2006).

Data mining technology provides a user oriented approach to explore hidden patterns in the data. The discovered knowledge/patterns can be used by the health care administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives (Kaur and Wasan, 2006).

Low birth weight (LBW) is the main factor determining neonatal and prenatal survival and it is also associated with many adverse outcomes in newborns. A baby's low weight at birth is either the result of preterm birth (before 37 weeks of gestation) or of restricted foetal (intrauterine) growth (Kramer, 1987). He further described that low birth weight is closely associated with foetal and neonatal mortality and morbidity, inhibited growth and cognitive development, and chronic diseases later in life.

Many factors affect the duration of gestation and of foetal growth, and thus, the birth weight. They relate to the infant, the mother or the physical environment and play an important role in determining the birth weight and future health of the infant (WHO Technical Consultation, 2004). The studies show that birth weight is affected to a great extent by the mother's own foetal growth and her diet from birth to pregnancy, and her body composition at conception. Mothers in poor socio-economic conditions frequently have low birth weight infants. In those settings, the infant's low birth weight stems primarily from the mother's poor nutrition and health over a long period of time, including during pregnancy, the high prevalence of specific and non specific infections, or from pregnancy complications underpinned by poverty. Physically demanding work during pregnancy also contributes to poor foetal growth (WHO Technical Consultation, 2004; Kramer, 1987).

1.2. Statements of the Problem

More than 20 million infants worldwide, representing 15.5 percent of all births are born with low birth weight, 95.6 percent of them in developing countries. The level of low birth weight in developing countries (16.5 percent) is more than double the level in developed regions (7 percent) (UNICEF/WHO, 2004). The study shows that half of all low birth weight babies are born in South-central Asia, where more than a quarter (27

percent) of all infants weighs less than 2,500 grams at birth. Low birth weight levels in sub-Saharan Africa are around 15 percent. Central and South America have, on average, much lower rates (10 per cent), while in the Caribbean the level (14 percent) is almost as high as in sub-Saharan Africa (UNICEF/WHO, 2004).

In Ethiopia study have reported from singleton live birth hospital records, the incidence of low birth weight was around 11% in 1990's (Fikre Enkuslase, and Aklilu , 2000). Similar study conducted in Jimma zone reported that prevalence of low birth weight in that zone was 22.5%. Mothers residing in the urban setting had higher risk of delivering LBW babies and the difference was statistically significant. Analysis of maternal obstetric history revealed that those mothers who delivered before 37 weeks of gestation, had weight loss, and who did not receive additional diet during pregnancy had higher risk of delivering LBW babies. Similarly, those who had multiple gestations had a higher risk of delivering LBW babies (Tema, 2006).

However, all those previous studies were conducted by using a very small proportion of the datasets. Besides, in those studies, data analysis was conducted by using statistical techniques. Since the analysis made by using traditional 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. The other reason is Data Mining techniques tend to be more robust for real-world messy data and also used less by expert users (Plate, et al., 1997).

Thus the goal of this study is to predict low birth weight using EDHS 2005 (Ethiopia Demographic and Health Survey) data set by applying of data mining techniques. This study is tried to build a model and addressed the factors associated with low birth weight using EDHS data set by applying data mining Techniques.

1.3. Objective of the Research

1.3.1 General Objectives

The general objective of this study is to identify the determinate factors of low birth weight and to develop a predictive model for low birth weight that can support health practitioner in the sector.

1.3.2 Specific Objectives

The specific objectives are:

- To review empirical as well as conceptual literatures to understand problems associated with low birth weight and the use of data mining.
- To collect, prepare and pre-process the raw data for model building by selecting, cleaning, and constructing the birth weight data set.
- To develop predictive model that help to identify LBW determinant factors.
- To compare the models based on the classifier evaluating criteria.
- To identify the determinant factors of low birth weight.
- To report result and forward recommendation.

1.4 Methodology

1.4.1 Study Area

The study is set in both urban and rural areas of Ethiopia. A total of 11 geographic areas (9 regions and 2 city administrations), namely: Tigray; Affar; Amhara; Oromiya; Somali; Benishangul-Gumuz; Southern Nations, Nationalities and Peoples (SNNP); Gambela; Harari; Addis Ababa and Dire Dawa are included in the study.

1.4.2 Data collection

The research is carried out based on secondary data, collected from (EDHS) Ethiopia Demographic Health and Survey data set from MEASURE DHS (http://www.measuredhs.com), which is available for researchers. Nationally, surveys collect a wealth of information on widely different topics in a specific country. MEASURE DHS supports a range of data collection options that can be customized to fit

specific monitoring and evaluation needs of host countries. The obtained original EDHS data sets are in SPSS (Statistics Package for Social Sciences) format.

1.4.3 Study design

In order to achieve the above stated objectives, the researcher has used the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model, which contains six phases as shown in Figure 1.1.

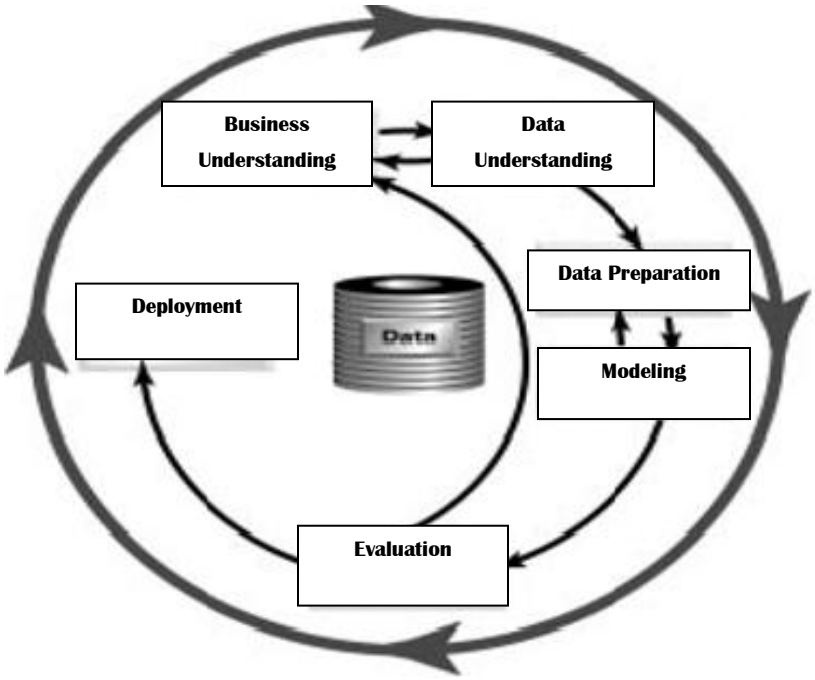

Figure 1.1 Phases of the CRISP-DM reference model

Source: Chapman et al, (2000)

Business understanding- this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives. In this study in order to understand the business and applied the above steps, work closely with a domain experts and review of related literatures are performed.

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

The primary data source for this study is the 2005 EDHS data sets, which contains details information on infant, child, adult and maternal mortality, fertility, maternal and child health etc. Hence to get familiar with the data, to identify data quality problems, to have insights into the data and to detect interesting subsets of the data to be used different literatures have been surveyed. Besides, careful analysis of the data and its structure has been done together with the domain experts.

Data preparation- Once the data resources available are identified, the dataset is prepared by cleaning the missing values with appropriate methods i.e numerical value replaced with mean and nominal values are replaced with modal value. Because Cleaning and data transformation, preparation of data set for modeling is needs to occur in this phase. Data exploration at a greater depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding.

Modeling: - In this phase various modeling techniques are selected and applied, and their parameters are adjusted to optimal values. Although the choice of data mining techniques for classification tasks seems to be strongly dependent on the application, one of the data mining techniques that are frequently employed for classification tasks is decision tree. As it is indicated previously, the purpose of this research is to develop predictive model. Decision tree classification data mining technique is applied in predicting low birth weight on EDHS data set. To this end, J48 classification decision tree and rule induction techniques to the problem domain is employed and tested.

Evaluation-at this stage, the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives. Evaluation is the key to making real progress in data mining. After building a model, we must evaluate its results and interpret their significance (Two Crows Corporation, 2005). In order to evaluate the performance of a model before deployment, there is a need to examine the error rate on the data set that did not take part in the process of model formulation.

In this thesis, the analysis of error rate generated by the confusion matrix is used to compute accuracy which measures the result of classification. The other evaluating criteria used for the models performance evaluation is the detailed accuracy measure, which measures the true positive rate, the false positive rate, and the precision and recall of the models developed. Accordingly a model with high success rate or low error rate having high precision, higher ROC (Receiver Operating Characteristics) curve and recall are considered as a good model. And this is followed by subject matters expert consultation for validity and acceptance of the models in the study domain.

Deployment-creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The sequence of the phases is not rigid. Moving back and forth between different phases is usually required and possible (Chapman et al, 2000).

1.4.4 Sampling

In the process of developing a predictive model, it is necessary to prepare well defined training and testing dataset to assure that the model is the most accurate prediction. In this study tenfold cross validation random sample generation technique has been used, in setting the training and testing data set samples, where training dataset is used to train and build the models and test dataset are used to test the performance of the model. Studies showed that ten seem to be an optimal number of folds that optimizes the time it takes to complete the test and the bias and variance associated with the validation process (Witten and Frank, 2005).

1.4.5 Data Mining Tool Selection

Many good data mining software products are being used, ranging from well-established, Enterprise Miner by SAS and Intelligent Miner by IBM, CLEMENTINE by SPSS, PolyAnalyst by Megaputer, and many others in a growing and dynamic industry. WEKA (from the University of Waikato in New Zealand) is an open source tool with many useful machine learning methods (David and Delen, 2008).

Data mining tools need to be versatile, scalable, capable of accurately predicting responses between actions and results, and capable of automatic implementation (Chackrabarti, et al, 2009). Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to the business strategies, knowledge bases, scientific and medical research (Han and Kamber, 2006).

Data mining tools predict future trends and behaviors and help organizations to make practical knowledge-driven decisions (Larose, 2005). The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were more time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations (Deshpande and Thakare, 2010).

In this research, Weka 3.6.3 software is used as a mining tool. In addition Microsoft Excel for data cleaning and for converting the original file to CSV file format, and Microsoft Word for documentation purpose have been used.

There are factors that contribute to the usefulness of data mining tools or software to the intended data mining tasks: The tool selected should be able to provide the required data mining functions. The data mining functionality that the researcher has intended to carry out in this research is prediction. In addition the methodologies used by the data mining software to perform each of the data mining functions are also important factor to consider. The researcher has chosen the J48 decision tree classifier implementation that implements C4.5 algorithm.

In addition, the selected tool can comfortably operate on windows operating system and stand alone environment. Hence Windows 7 operating system on a standalone machine has been utilized. Another important consideration in tool selection is visualization capabilities. The variety, quality and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of data mining systems. Weka has a facility to visualize its output in this regard (Witten and Frank, 2005).

The other reason behind the selection of Weka for this study is familiarity of the researcher with the tool, its comprehensiveness for this study requirements and the ease of availability of the tool; Weka provides a number of data mining functionalities, Such as classification, clustering, association, attribute selection and visualization. Weka is developed at the University of Waikato, in New Zealand. 'Weka' stands for the Waikato Environment, Knowledge Analysis. The system is written in JAVA an object oriented programming language, and has been tested under Linux, windows and Macintosh operating systems. Weka includes varieties of tools, for preprocessing a data set, such as attribute selection, attribute filtering and attribute transformation (Witten and Frank, 2005).

1.5 Scope and limitations of the Research

According to measure Ethiopian the last survey in Ethiopia is 2005 EDHS, because the survey is conducted within five years intervals. The 2010 EDHS data set may finalize at the end of august 2011. Therefore, the scope of this research is limited to develop a predictive model for low birth weight using the 2005 Ethiopia Demographic and Health survey data sets. Moreover the study was limited to the development of predictive model by using CRISP-DM process models, and does not incorporate the deployment phase of the model. Due time and the deployment phase is need more the involvement of users and domain experts.

Due to the nature of the available dataset and the time available to complete the study, the researcher has limited the study to predict only Low birth weight. Such as the determinant factors of preterm birth is not identified using the prediction model developed by this study.

1.6 Significance of the Research

Data mining technology provides a user oriented approach to novel and hidden patterns in the data. The discovered patterns (knowledge) can be used by the health care administrators to improve the quality of service. Predicting low birth weight and extracting the knowledge from birth weight data is very helpful to reducing child mortality.

The primary goal of this research is to build a predictive model by using data mining techniques. The predicting of low birth weight among live birth babies would be based upon demographic, parental, and epidemiological history factors, without using diagnostic tests or physical exam findings. The built models and rules can support the national health care policy in revising the existing rules, and inducing new rules and policies. It is also possible to use the research result, the predictive model, as a framework for improving maternal health care. This type of model building might have an application outside of clinical settings to support health care workers in taking preventive actions to reduce low birth weight in the country, as well as to assist health care planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at improving child survival in the regions as well as in the country.

1.7. Thesis Organization

The structure of this thesis is presented as follows. Chapter one provides details information about background of the study, statements of the problem, objectives of the study, methodology, scope and limitation of the study, and significance of the study.

The second chapter also provides details information about conceptual and related literature review of data mining technology and low birth weight respectively. Details description of data set, data preparation and preprocessing activities are described in chapter three.

Chapter four is deals with conducting different experiments using J48 decision tree and PART rule induction algorithms by giving the parameters different values. Both J48 decision tree and PART rule induction algorithms used all and reduced attributes to conduct the experiments. In addition, it presents results of the experiments and their interpretations. Chapter five, the thesis is concluded in this chapter, and details of future work in this area and recommendations are presented.

CHAPTER TWO

LITERATURE REVIEW

2.1 Over View of Data Mining

Technology now allows us to capture and store vast quantities of data. The amount of data in the world, in our lives, seems to be increasing and there's no end in sight (Witten and Frank, 2005). Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amount of data. We are overwhelmed by data - medical data, demographic data, financial data and marketing data (Han and Kamber, 2006). It has been estimated that the amount of data stored in the world's databases doubles every 20 months (Witten and Frank, 2005).

To undertake large data analysis project, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems (Hand, Heikki and Smyth, 2001).

Witten and Frank (2005) states lack of data is no longer a problem at the current stage. However the inability to generate useful information from data is the problem. As the volume of data increases inexorably, the proportion of people understands decreases, alarmingly. Lying in hidden data, in all these data potentially useful information, i.e. rarely made explicit or taken advantage of.

According to Kumar et al (2008) the health care environment is generally perceived as being 'rich in information' yet having 'knowledge poor'. There is a wealth of data available within the health care systems. Baylis (1999) states health care generates large amounts of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce large amounts of clinical data. This data is a strategic resource for health care institutions.

The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making (Koh and Tan, 2006). The explosive growth in raw data volume generates the need for new techniques and tools that can intelligently and automatically transform the data into useful information and knowledge.

Today, the size of the world population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did. This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector (Ruben and Canlas, 2009). Witten and Frank (2005) has states that, as the world grows in complexity over whelming us with data it generates, data mining becomes our only hope for elucidating the patterns that underlie it. Intelligently analyzed data is a valuable resource. It can lead to new insights and, in commercial settings, to competitive advantages.

2.2. Data Mining

Data can be form simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of these data; data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining' (Deshpande and Thakare, 2010).

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand, Mannila and Padhraic, 2001). Data mining technologies have proven to be useful and effective in different areas, including marketing, customer

relationship management, engineering, medical and biomedical research (Fayyad, et al. 1996).

Data mining, popularly known as Knowledge Discovery in Databases (KDD) is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases (Dunham and Sridhar, 2006). Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process (Fayyad, et al. 1996; Bramer. 2007). Han and Kamber (2006), defined data mining, as the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

According to Berry and Linoff (2004), data mining usually makes sense when there is large amount of data. For this reason most of the algorithms developed for data mining purpose requires large volume of data so as to build and train models that are responsible for different tasks of data mining such as classification, clustering, prediction, association and the like. Moreover, the need for bulky data can be explained by a couple of reasons. Primarily in the case of small databases, it is feasible to capture appealing trends and relationships by introducing traditional tools such as spreadsheets and database query. The second reason is that most data mining tools and algorithms demand large amount of training data (data used for building a model) in order to generate unbiased models. The rationale is simple and straight forward, small training data results in unreliable generalizations based on chance patterns.

2.3 Data Mining or Knowledge Discovery Process

Data mining or knowledge discovery refers to the process of finding interesting information in large repositories of data. The term data mining also refers to the step in the knowledge discovery process in which special algorithms are employed in hopes of identifying interesting patterns in the data. These interesting patterns are then analyzed yielding knowledge (Bowen, 2006).

Discovering knowledge in data presents data mining as a well-structured standard process, intimately connected with managers, decision makers, and those involved in deploying the results (Larose, 2005). Therefore both novices and data-mining specialists need assistance in knowledge discovery processes (Deshpande and Thakare, 2010). A cross-industry standard was clearly required that is industry neutral, tool-neutral, and application-neutral (Larose, 2005).

This section describes CRISP-DM (CRoss-Industry Standard Process for Data Mining), a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit. Developed by industry leaders with input from more than 200 data mining users and data mining tool and service providers, CRISP-DM is an industry-, tool-, and application-neutral model. This model encourages best practices and offers organizations the structure needed to realize better, faster results from data mining. The CRISP–DM demands that data mining be seen as an entire process, from communication of the business problem through data collection and management, data preprocessing, model building, model evaluation, and finally, model deployment (Chapman et al, 2000). Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it (Chapman et al, 2000).

According to Chapman et al (2000), a given data mining project has a life cycle consisting of six phases. Note that the phase sequence is adaptive. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase (Larose, 2005).

CRISP-DM is complete and well documented. All the stages are properly organized, structured and defined, allowing that a project could be easily understood or revised (Santos & Azevedo, 2005).

2.3.1 CRISP–DM: The Six Phases

2.3.1.1 Business (Problem) understanding phase

The first phase in the CRISP–DM standard process may also be termed the research understanding phase. According to Shearer (2000) the most important phase of any data mining project is the initial business understanding phase which focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed and how, it is vital for data mining practitioners to fully understand the business for which they are finding a solution.

The business understanding phase involves several key steps; including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan (Chapman et al, 2000). The two crows' corporation (2005) also stated that to make the best use of data mining one must make a clear statement of the objective. Without clear understanding of the problem to be solved, it is difficult to identify, select, and prepare the data for mining or to correctly interpret the results.

2.3.1.2 Data understanding

According to Chapman et al (2000) the data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

2.3.1.3 Data Preparation Phase

The data preparation and preprocessing phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the

cleansing of data, and the construction of data, the integration of data, and the formatting of data (Chapman et al, 2000; Larose, 2005).

This phase is often the most time consuming task of KDD processes, especially if data is drawn directly from the company's operational databases rather than from the data warehouse. As mentioned by Han and Kamber (2006) the data stored in databases may reflect noise, exceptional cases, or incomplete data objects. When mining data regularities, these may confuse the process, causing the knowledge model constructed to over fit the data. As a result the accuracy of the discovered patterns can decrease.

2.3.1.4 Building Model Phase

In this phase, various modeling techniques are selected and applied and their parameters are adjusted to optimal values. Typically, several techniques exist for the same data mining problem type. Two crows' corporation (2005) describes that the most important thing to remember about model building is that it is an iterative process. You will need to explore alternative models to find the one that is most useful in solving your business problem. What you learn in searching for a good model may lead you to go back and make some changes to the data you are using or even modify your problem statement. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models (Chapman et al, 2000; Shearer, 2000).

2.3.1.5 Evaluation Phase

Before proceeding to the final deployment of the model, it is important to thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. Here it is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps (Shearer, 2000).

2.3.1.6 Deployment Phase

Model creation is generally not the end of the project. The knowledge gained must be organized and presented in a way that the customer can use it, which often involves applying "live" models within an organization's decision-making processes, such as the real-time personalization of Web pages or repeated scoring of marketing databases (Shearer, 2000).

According to Chapman et al (2000), depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. Even though it is often the customer, not the data analyst, who carries out the deployment steps, it is important for the customer to understand up front what actions must be taken in order to actually make use of the created models.

2.4 Data mining Tasks

Data mining functionalities 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 (Han and Kamber, 2006).

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined (Deshpande and Thakare, 2010). 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 (Han and Kamber, 2006).

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. Where as descriptive techniques are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms (Two Crows Corporations, 2005).

According to Fayyad, et al (1996) the goals of prediction and description can be achieved using a variety of particular data-mining methods. The data mining methods are broadly categorized as: On-Line Analytical Processing (OLAP), Classification, Clustering, and Association Rule mining, etc. These methods use different types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. (Deshpande and Thakare, 2010). The Data mining methods that are used for this study and the behavior of the patterns it discovers is described below.

2.4.1 Classification and Prediction

Classification, one of the most common data mining methods, seems to be a human imperative. Human beings usually classify, categorize or grading in order to understand and communicate about the world (Berry and Linoff, 2004). Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes. It is a supervise learning because the classes are predefined before the examination of the target data (Han and Kamber, 2006; Two crows corporations, 2005). Pang-Ning et al (2006) also stated that classification is the task of learning a target function f that maps each attribute X to one of the predefined class label Y.

According to Two Crows Corporations (2005), classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Han and Kamber (2001) describing the classification task stated that data classification is a two-step process. In the first step, a model is constructed by analyzing database tuples

described by the attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. In the second step the model is used for classification. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. Then the accuracy of a model on a given test set is evaluated.

Prediction is the same as classification, except that the records are classified according to some predicted future behavior or estimated future value. The primary reason for treating prediction as a separate task from classification and estimation is that in predictive modeling there are additional issues regarding the temporal relationship of the input variables or predictors to the target variable.

Any of the techniques used for classification can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior (Berry and Linoff, 2004).

For this research, a classification task is to be carried out since a model is to be built by using the pre-classified data of past records, the 2005 EDHS birth data in Ethiopia. Although the selection of techniques suitable for classification is mainly depends on the type of data used for mining and the expected outcome of the mining process. The domain experts play a significant role in the selection of algorithm for data mining (Deshpande and Thakare, 2010)

2.4.2 Issues in classification and prediction

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

2009). Most classification algorithms have some mechanisms for handling noisy or missing data; this step can help to reduce confusion during learning (Han and Kamber, 2005).

One of the most important things to do, before starting the data mining process, is to make sure your data is as clean as possible. This is probably the most tedious, and time consuming part of data mining. But in order to get the best results possible, the data must be in good form, so to speak. Data cleansing involves making sure the values of attributes are consistent and normalized (Moheb , et al, 2005). Therefore organizations are forced to think about a unified logical view of the wide variety of data and databases they possess, they have to address the issues of mapping data to a single naming convention uniformly representing and handling missing data, and handling noise and errors when possible (Fayyad, et al. 1996).

Relevance analysis: According to Han and Kamber (2005) many of the attributes in the data may be redundant. Correlation analysis can be used to identify whether any two given attributes are statistically related. For example, a strong correlation between attributes A1 and A2 would suggest that one of the two could be removed from further analysis. A database may also contain irrelevant attributes. Attribute subset selection can be used in these cases to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Hence, relevance analysis, in the form of correlation analysis and attribute subset selection, can be used to detect attributes that do not contribute to the classification or prediction task. Including such attributes may otherwise slow down, and possibly mislead, the learning step.

Data transformation and reduction: - Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as -1:0 to 1:0, or 0:0 to 1:0. In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (e.g. income) from outweighing attributes with initially smaller ranges (such as binary attributes). Data reduction can reduce the

data size by aggregating, eliminating redundant features, or clustering, for instance. These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

The process of data transformation might include smoothing (e.g. using bin means to replace data errors), aggregation (e.g. viewing monthly data rather than daily), generalization (e.g. defining people as young, middle-aged, or old instead of by their exact age), normalization (scaling the data inside a fixed range), and attribute construction (adding new attributes to the data set) (Chakrabarti et al., 2009; Han and Kamber, 2006).

2.5 Commonly used classification techniques in data mining

Understanding and learning of different data mining algorithms for classifications and prediction is essential, in order to take advantage of specific classification and prediction techniques , to decide the best and appropriate technique for the problem at hand and to know the advantages and disadvantages of a technique.

Classification is one of the most frequently studied problems by Data Mining and machine learning researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). The most commonly used data mining classification methods are:

- Decision tree,
- Rule induction,
- Neural Network,
- Genetic algorithm and
- Nearest neighbor method (Romero C., et al, 2008).

Knowledge is usually represented in the form of rules indicating the degree of association between two variables, rules mapping data into predefined classes, rules that identify a finite set of categories or clusters to describe the data, etc. These rules support specific tasks and are generated by repeated application of a certain technique, or more generally an algorithm, on the data. The quality of these rules and hence the knowledge discovered

is heavily dependent on the algorithms used to analyze the data. Thus, central to the problem of knowledge extraction are the techniques/methods used to generate such rules (Deogun , et al, 2001). Two of the methods that are used in this thesis are discussed in following section.

2.5.1 Rule induction

Rule induction is one of the major forms of data mining and is perhaps the most common form of knowledge discovery in unsupervised learning systems. It is also perhaps the form of data mining that most closely resembles the process that most people think of when they think about data mining, namely "mining" for gold through a vast database. The gold in this case would be a rule that is interesting - that tells you something about your database that you didn't already know and probably weren't able to explicitly articulate (aside from saying "show me things that are interesting") (Thearling, et al, 2010).

Rule induction is the process of extracting useful 'if then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form

IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute (Srinivas et al, 2010). He further explained an accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) ----> (Cause—of--- disease)

Example 1: If_then_rule induced in the diagnosis of level of alcohol in blood:

IF Sex = MALE
AND Unit = 8.9
AND Meal = FULL
THEN Diagnosis=Blood_alcohol_content_HIGH.

When the rules are mined out of the database the rules can be used either for better understanding of the business problems that the data reflects or for performing actual predictions against some predefined prediction target and it is help full for decision making in healthcare(Srinivas et al , 2010; Kaur and Kirishan, 2006).

2.6 Decision tree

A decision tree is a technique that produces a graphical analysis of the model it produces. The graphic output consists of a tree with nodes denoting decision points. The decision tree method encompasses a number of specific algorithms including Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 and C5.0. A decision tree is a model that is both predictive and descriptive. Alberto (2000) describes that decision trees are commonly used for classification but can also be used for regression analysis. Decision trees are advantageous tools for making corporate or financial decisions where a lot of complex information has to be considered. Decision trees provide a functional framework in which alternate decisions and the implications of making those decisions can be laid down and evaluated. Decision trees also help in forming a balanced, accurate picture of the risks and rewards that can result from a particular decision.

Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database systems, and they have comparable or better accuracy in many applications (Kaur and Kirishan, 2006). Decision trees are considered easily understood models because a reasoning process can be given for each conclusion. However, if the tree obtained is very large (a lot of nodes and leaves) then they are less comprehensible. A decision tree can be directly transformed into a set of IF-THEN rules that are one of the most popular forms of knowledge representation, due to their simplicity and comprehensibility (Romero, et al, 2008).

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 (Larose, 2005).

2.6.1 Basic Decision Tree Construction Algorithm

There are many algorithms for growing a decision tree. Most of them have a core mechanism that employs a top-down, greedy construction of the decision tree. The ID3 algorithm is a good example of decision tree construction using a top-down approach. It begins by determining which feature should be tested at the root of the tree. This is done by evaluating each feature using a statistical test to examine how well it alone classifies the training samples. The best feature is selected to be tested at the root node of the tree (Phyu, 2009).

As demonstrated in Figure 2.2 Each non-leaf node is connected to a test that splits it's set of possible answers into subsets corresponding to different test results. Each branch carries a particular test result's subset to another node. Each node is connected to a set of possible answers (Phyu, 2009). It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute) (Kaur and Kirishan, 2006).

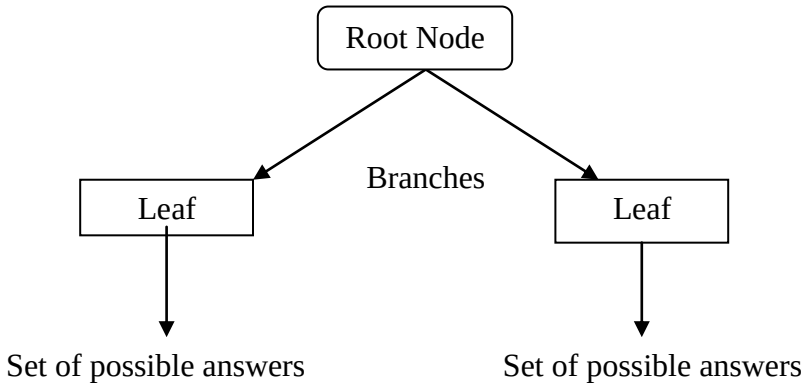

Figure 2.2 Example of decision tree structure

Decision trees are turned upside down and built from the root at the top toward the leaves at the bottom. The nodes of a tree represent questions; an answer to one question determines which question will be asked next. The process starts in the root node, where a record is tested and the result of the test determines lower node where the process will

proceed. It is an iterative process that is repeated until the record reaches a leaf, which represents one class of the data. Every node in the tree represents a test of some case attribute, and a path that leads from root to the leaf represents a rule that was used for classification, that is, every branch that is derived from that node represents a possible value of that attribute.

Various algorithms are used to choose tests based on how well the tests separate target classes. Choosing a different set of tests, or even a different sequence of the same set, results in a different tree are not the same. That means when a test is chosen and subset partitioning is made, most of the algorithms do not go back and question alternative possibilities as a means to choose the simplest (smallest) tree. This is a property of "greedy algorithms"; most of the algorithms for building decision trees fall into this category.

Decision trees can be binary, ternary, etc. depending on a count of different answers. A binary tree is the one that, for instance, answers the question with „yes" or „no", so that every leaf has two „child" nodes, and the answer determines which way the data will go to the next level. If the data has m attributes, the maximum height of a tree will be m.

We can measure effectiveness of a tree as a whole with its application to new data and by viewing the percentage of the data that is correctly classified. In every node, we can measure: a number of records that enter the node, the way these records will be classified if that was the leaf node, the percentage of records that are correctly classified in that node (Berry and Linoff, 2004 ; Mirjana and Dijana , 2008).

2.6.2 Attribute Selection Measure

Most Decision Tree classifiers perform classification in two phases: *tree-induction (growing or building*) and *tree-pruning*. In the tree-induction phase the algorithm starts with the whole data set at the root node. The data set is partitioned according to a splitting criterion into subsets. This procedure is repeated recursively for each subset until each subset contains only members belonging to the same class or is sufficiently small. In the

tree-pruning phase the full grown tree is cut back to prevent over-fitting and to improve the accuracy of the tree (Aurelian, 2007).

During the induction phase of the Decision Tree, the attribute selection measure is determined by choosing the attribute that will best separate the remaining samples of the nodes partition into individual classes. A critical problem in building decision trees is the attribute selection measure problem. Attribute selection is the process of including/removing certain attributes or columns of data for analysis (Moheb, et al, 2005; Liangxiao and Chaoqun , 2010).

The problem of building a decision tree can be expressed recursively. First, a best attribute is selected to place at the root node of the tree and create one child node for each possible value of this attribute. For each child node, if it isn't a leaf node, the entire process is then repeated recursively only using those training instances that actually reach this node. If it is a leaf node, stop splitting this branch of the tree. Obviously, two critical problems must be addressed in building a decision tree. One is which attribute is the best, and the other is how to judge a node is a leaf node (Liangxiao and Chaoqun , 2010). Most machine learning algorithms are designed to learn which attributes are the most appropriate to use for making their decisions (Witten and Frank, 2005).

As stated, the best feature under some criterion is chosen as the test at the root node, and later other features are chosen in the same way as roots of sub trees. Several optional criteria can be used. The ID3 and C4.5 algorithm uses the information gain measure, which computes how well a given feature separates the training samples according to their class labels. To define it, we first introduce the entropy measure from information theory (Quinlan, 1986).

$$Entropy(S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where c is the number of different classes (labels), and p i is the proportion of S (the group of samples) belonging to class i.

Notice that the entropy is 0 if all members of S belong to the same class. If all the classes contain an equal number of samples ($p_i = \frac{1}{c}$ for all i)then the entropy of S is equal to 2

log c , which equals the minimum number of bits needed to encode the classification of an arbitrary sample in S, when c is a power of 2. In the specific case where c=2, if both classes have the same number of samples then the entropy of S is equal to 1. This way, entropy gives us a measure of the impurity of the sample group.

Now we can define the information gain measure. It is simply the expected reduction in entropy caused by partitioning the samples according to a particular feature. The information gain, Gain*(S, F)* of a feature F, given a collection of samples S, is defined as (Quinlan, 1986)

$$Gain(S,F) \equiv Entropy(S) - \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where Values (F) is the set of all possible values for feature F, and v is the subset of S consisting of samples for which feature F has the value v. Hence, Gain(S, F) is the information provided (the reduction in entropy) about the target function value (the class label); given the values of a particular feature F (Quinlan, 1986).

2.6.3 Avoiding Model Over-fitting

Decision tree classifiers aim to refine the training sample into subsets which have only a single class. However, training samples may not be representative of the population they are intended to represent. In most cases, fitting a decision tree until all leaves contain data for a single class causes *over-fitting*. That is, the decision tree is designed to classify the training sample rather than the overall population and accuracy on the overall population will be much lower than the accuracy on the training sample (Podgorelec, et al., 2002).

Most of the decision tree induction algorithms (ID3, C4.5, CART…) use pruning. They all grow trees to maximum size, where each leaf contains single-class data or no test offers any improvement on the mix of classes at that leaf. There are several approaches to avoid over-fitting, and they can be grouped into two classes. The first class of approaches stop growing the tree before it reaches its full potential size. The second class of approaches fully grow the trees (perhaps causing over-fit of the data), and then prune it. Pruning it means removing the sub-tree rooted at that node, thus making it a leaf node. The assigned class-label of this node is the majority class in the training samples

associated with that node. In this approach, nodes are removed when the resulting pruned tree performs no worse than the original tree over the validation set (Podgorelec, et al., 2002; Gutkin, 2008). There are two approaches in tree pruning.

Pre-pruning:- a tree is pruned by halting its construction early. (E.g. by deciding not to further split or partition the subset of training samples at a given node.) In this approach the tree growing algorithm is halted before generating fully grown tree that perfectly fit the entire training data. To do this a more restrictive stopping condition must be used. E.g. stop expanding a leaf node when the observed gain in impurity measure falls below certain threshold. Nevertheless it is difficult to choose the right threshold for early termination. If threshold too high it results in under fitted model, if too low, it may not be sufficient to overcome the model over fitting problem (Ping_Ning, et. al, 2006).

Post pruning: - removes branches from fully grown tree. A tree node is pruned by removing its branches i.e. the decision tree is first grown to its maximum size. This is followed by a tree pruning step, which proceeds to trim the fully grown tree in a bottom up fashion. Trim can be done by replacing a sub tree with a new leaf node whose class label is determined from the majority class records affiliated with the sub tree. Post pruning tends to give better results than pre pruning because it makes pruning decisions based on a fully grown tree, unlike pre pruning which can suffer from pre mature termination of the growing process (Ping_Ning, et. al, 2006).

2.7 Classifier Accuracy (performance evaluation) Measures

The objective of a model obtained through the data-mining process is to classify/predict new instances correctly. The commonly used measure of a model's quality is predictive accuracy. Since new instances are supposed not to be seen by the model in its learning phase, we need to estimate its predictive accuracy using the true error rate. The true error rate is statistically defined as the error rate of the model on an asymptotically large number of new cases, where this number converges to the actual population distribution. In practice, the true error rate of a data-mining model must be estimated from all the available samples, which are usually split into training and testing sets. This prevents the problem of over fitting and gives a better measure of the accuracy of the generated models (Kantardzic , 2003).

According to David and Delen (2008), estimating the accuracy of a classifier induced by some supervised learning algorithms is important for the following reasons. First, it can be used to estimate its future prediction accuracy which could imply the level of confidence one should have in the classifier's output in the prediction system. Second, it can be used for choosing a classifier from a given set (selecting the "best" model from two or more qualification models). Lastly, it can be used to assign confidence levels to multiple classifiers so that the outcome of a combining classifier can be optimized. Combined classifiers are increasingly becoming more popular due to the empirical results that suggest them producing more robust and more accurate predictions as they are compared to the individual predictors. For estimating the final accuracy of a classifier one would like an estimation method with low bias and low variance. In some application domains, to choose a classifier or to combine classifiers the absolute accuracies may be less important and one might be willing to trade off bias for low variance. Some of the most popular estimation methodologies used for classification type data mining models are discussed bellow.

Simple Split (Holdout):- The simple split (holdout or test sample estimation) partitions the data into two mutually exclusive subsets called a training set and a test set (or holdout set). It is common to designate 2/3 of the data as the training set and the remaining 1/3 as the test set. The training set is used by the inducer (model builder) and the built classifier is then tested on the test set. The processed data is partition into three mutually exclusive subsets; training, validation, testing. The validation set is used during the model building to prevent the over-fitting (David and Dursun, 2008; Han and Kamber, 2006).

They stated further, the main criticism of this method is the fact that it makes the assumption that the data in the two subset are of the same kind (has the exact same properties). Since this is a simple random partitioning, in most realistic datasets where the data is skewed on the classification variable, such an assumption may not hold true. In order to improve this situation, stratified sampling is suggested, where the strata becomes the output variable. Even though this is an improvement over the simple split, it still has a bias associated from the single random partitioning.

The k-Fold Cross Validation:-In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, one can use a methodology called k-fold cross validation. In k-fold cross validation, also called rotation estimation, the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. The classification model is trained and tested k times. Each time it is trained on all but one folds and tested on the remaining single fold (David and Dursun , 2008).

The k-fold cross validation is also called 10-fold cross validation, because the k taking the value of 10 has been the most common practice. In fact, empirical studies showed that ten seem to be an optimal number of folds (that optimizes the time it takes to complete the test and the bias and variance associated with the validation process) (Witten and Frank, 2005). 10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation. In fact, in data mining community, for methods-comparison studies with relatively smaller datasets, k-fold type of experimentation methods are recommended. In essence, the main advantage of 10-fold (or any number of folds) cross validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as holdout sample (David and Dursun , 2008). In this study 10-fold cross validation will be used because of the above reasons.

Bootstrapping:-Bootstrapping is another technique for estimating the error of a model; it is primarily used with very small data sets. As in cross validation, the model is built on the entire dataset. Then numerous data sets called bootstrap samples are created by sampling from the original data set. After each case is sampled, it is replaced and a case is selected again until the entire bootstrap sample is created. Note that records may occur more than once in the data sets thus created. A model is built on this data set, and its error rate is calculated. This is called the re-substitution error. Many bootstrap samples (sometimes over 1,000) are created. The final error estimate for the model built on the whole data set is calculated by taking the average of the estimates from each of the bootstrap samples (Two Crows Corporation, 2005).

Based upon the results of model building, it is possible to build another model using the same technique but different parameters, or perhaps try other algorithms or tools. For example, another approach may increase accuracy. No tool or technique is perfect for all data, and it is difficult if not impossible to be sure beforehand which technique will work the best. It is quite common to build numerous models before finding a satisfactory one (Two Crows Corporation, 2005).

Confusion matrix: - A confusion matrix is a very useful tool for understanding results; Confusion matrix shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong.

Given two-class, a confusion matrix may be used to summarize the predictive performance of a classifier on test data. It is commonly encountered in a two-class format, but can be generated for any number of classes. Suppose we have a two-class problem with classes referred to as positive and negative. A single prediction by a classifier can have four outcomes which are displayed in the confusion matrix Table 2.1.

Table: 2.1 Two-class confusion matrixes (also known as a 2 × 2 contingency table).

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True positive	False negative
	Negative	False Positive	True negative

True positive: the actual class of the test instance is positive and the classifier correctly predicts the class as positive.

False negative: the actual class of the test instance is positive but the classifier incorrectly predicts the class as negative.

False positive: the actual class of the test instance is negative but the classifier incorrectly predicts the class as positive.

True negative: the actual class of the test instance is negative and the classifier correctly predicts the class as negative.

For a particular data set the confusion matrix will contain in its cells the number of instances for each of the four possible classification outcomes. Note that the "correct" predictions lie on the main diagonal of the matrix while the "incorrect" predictions are in the off-diagonal cells (Han and Kamber, 2006).

2.8 Low Birth Weight

As described in chapter one in details Low birth weight refers to infants who weigh less than 2,500 grams at birth. Most normal babies weigh 2500 g by 37 weeks of gestation (WHO Report, 1992).

According to UNICEF/WHO (2004) reports More than 20 million infants worldwide, representing 15.5 per cent of all births, are born with low birth weight, 95.6 per cent of them in developing countries. The level of low birth weight in developing countries (16.5 per cent) is more than double the level in developed regions (7 per cent). Half of all low birth weight babies are born in South-central Asia, where more than quarters (27 per cent) of all infants weigh less than 2,500 grams at birth. Low birth weight levels in sub-Saharan Africa are around 15 per cent. Central and South America have, on average, much lower rates (10 per cent), while in the Caribbean the level (14 per cent) is almost as high as in sub-Saharan Africa. About 10 per cent of births in Oceania are low birth weight births.

2.8.1 Determinants of low birth weight

A baby's low weight at birth is either the result of preterm birth (before 37 weeks of gestation) or of restricted foetal (intrauterine) growth. Low birth weight is closely associated with foetal and neonatal mortality and morbidity, inhibited growth and cognitive development, and chronic diseases later in life. Many factors affect the duration of gestation and of foetal growth, and thus, the birth weight. They relate to the infant, the mother or the physical environment and play an important role in determining the infant's birth weight and future health (WHO Report, 1992; WHO Technical Consultation, 2006)

It is generally recognized that low birth weight can be caused by many factors. Factors with well-established direct causal impacts on intrauterine growth include infant sex,

racial/ethnic origin, maternal height, pre-pregnancy weight, paternal weight and height, maternal birth weight, parity, history of prior low-birth-weight infants, gestational weight gain and caloric intake, general morbidity and episodic illness, malaria, cigarette smoking, alcohol consumption, and tobacco chewing. In developing countries, the major determinants of intrauterine growth poor gestational nutrition, low pre-pregnancy weight, short maternal stature, and malaria. In developed countries, the most important single factor, by far, is cigarette smoking, followed by poor gestational nutrition and low pre-pregnancy weight. For gestational duration, only pre-pregnancy weight, prior history of prematurity or spontaneous abortion, and cigarette smoking have well established causal effects, and the majority of prematurity occurring in both developing and developed country settings remains unexplained (Kramer, 1987).

2.9 Application of data mining in healthcare

Health care generates large amount of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce mountains of clinical data. This data is a strategic resource for health care institutions (Baylis, 1999).

One of the biggest applications for the data mining tools is the medical healthcare industry. It used to take researchers many hours to collect information into books and matrixes to try to figure out the cause of disease. With data mining the medical healthcare organizations was able to:-

- Characterize patient behavior to predict office visits.
- Identify successful medical therapies for different illnesses.
- Analyze cause and effect of diseases.
- Detect disease outbreaks and preventable hospital deaths and
- Perform analysis of health care centers for better health policy-making,

(Moheb , et al., 2005; Ruben and Canlas, 2009)

The aim of this section intends to a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use today in medical research and public health. Health care organizations are implementing data mining technologies to

help control costs and improve the efficiency of patient care. Data mining can be used to help predict future patient behavior and to improve treatment programs. By identifying high-risk patients, clinicians can better manage the care of patients today (Rogers and Joyner, 2001).

Discovery of hidden patterns and relationships often goes unexploited. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established.

Researchers have presented an intelligent and effective heart attack prediction methods using data mining. Firstly, they have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weight age, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Five mining goals are defined based on business intelligence and data exploration. The goals are to be evaluated against the trained models. All these models could answer complex queries in predicting heart attack (Srinivas et al., 2010)

Harleen and Wasan (2006) studied the potential use of classification based data mining techniques such as Rule based, decision tree to massive volume of healthcare data. In particular they study using classification techniques on a medical data set of diabetic patients. Given patient records with corresponding diagnosis, data mining methods are able to diagnose new cases. For instance, in the domain of early diagnosis of diabetic nephropathy disease, the patient record of laboratory examination comprises of condition attribute.

Invnka et.al (2007) tried to identify possible associations between medicines used in pregnancy and preterm deliveries using data mining as a screening tool. They used data mining to identify possible correlates between preterm delivery and medicines used by 92,235 pregnant Danish women who took part in the Danish National Birth Cohort (DNBC). Then they evaluated the association between one of the identified exposures

(vaccination) and the risk for preterm birth by using logistic regression. The women were classified into groups according to their exposure to vaccination. The regression analyses were adjusted for the following covariates: parity, infant's gender, maternal Body-Mass Index (BMI), age, smoking, drinking, and job, number of inhabitants in the place of residence, infections, diabetes, high blood pressure and preeclampsia. Data mining had indicated that maternal vaccination (among other factors) might be related to preterm birth. Whether the association between maternal vaccination and the risk for preterm birth found here is causal or not deserves further studies. Data mining, especially with additional refinements, may be a valuable and very efficient tool to screen large databases for relevant information which can be used in clinical and public health research.

Another research conducted on women and their choice of contraceptive methods. The researcher use decision trees method to determine if there are common characteristics of the women and to their choice of contraception. It was that found the most important variable in case of women's choice of contraceptive methods is a husband's profession. The decision tree method is successfully applied in explanation of women's choice of contraceptive methods (Mirjana and Dijana , 2008).

CHAPTER THREE

DATA PREPARATION AND PRE-PROCESSING

3.1 Overview Data Pre-processing

It is well known that success of every data mining algorithm is strongly dependent on a quality of data processing. Data pre-processing task could be critical and a very complicated task. Sometimes, the data pre-processing takes more than half of the total time spent on the solving of the data mining problem, because incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases and data warehouses (Chackrabarti, et al., 2009; Ping-Ning, et al. , 2006). Thus, data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing 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 (Hu, 2003). It is an important step in data mining, because quality decisions must be based on quality data.

This section deals with the overviews of the data source, data cleaning and data transformation of the data employed in this study. In general the researcher has followed the steps of data mining process mentioned in the chapter one section 1.4. In this study the methodology adapted is CRISP-DM.

As it has been mentioned in section 1.3 the main objective of this study is to build a predictive model and develop rules for low birth weight babies. The successful predictive model can support health practitioner in Ethiopia.

Business understanding phase: The source of data for this research is the 2005 Ethiopian Demographic and Health Surveys census. The 2005 Ethiopia Demographic and Health Survey (EDHS) was conducted under the support of the Ministry of Health and implemented by the Population and Housing Census Commission Office (PHCCO), now merged with the Central Statistical Agency (CSA). The census is conducting in every five years intervals. The primary objective of the 2005 EDHS was to provide up-to-date information for policy makers, planners, researchers and programme managers, which

would allow guidance in the planning, implementation, monitoring and evaluation of population and health programmes in the country. The information obtained from the EDHS, in conjunction with statistical information obtained from the Welfare Monitoring Survey (WMS) and Household Income, Consumption and Expenditure Survey (HICES). It provides critical information for the monitoring and evaluation of the country's Plan for Accelerated and Sustained Development to End Poverty (PASDEP). The various sector development policies and programmes, and assist in the monitoring of the progress towards meeting the Millennium Development Goals (MDGs).

The 2005 EDHS collected information on the population and health situation which covers on family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, malaria, and women's empowerment

The 2005 Ethiopia Demographic and Health Survey is a nationally representative survey of 14,070 women age 15-49 and 6,033 men age 15-59. This sample provides estimates of health and demographic indicators at the national and regional levels, and for rural and urban areas. Among live births in the five years preceding the survey with a reported birth weight is **9861** records. Among all live births in the five years preceding the survey, the records contain mother's background characteristics. Mother's background characteristics are mother's age, region, and birth order number, place of residence, religion, wealth index, mother's education etc. On the other hand the senses include baby's information like sex of child, gestation week, weight of the baby and head size of the baby. The information related babies are collected from written record or mother's recall that are recorded during delivery. The original survey is kept in SPSS format. The selection and understanding of the dataset is performed by the help of domain expert.

After collecting the SPSS format original survey dataset, Microsoft Excel and Word is used for preparing the dataset into a form acceptable by the selected data mining software Weka.

3.2 Data Preparation

Data preparation is the most important phases of the data analysis activity which involves the construction of the final data set (data that will be fed into the modeling tool) from the initial raw data. Data preparation generates a dataset smaller than the original one, which can significantly improve the efficiency of data mining. This task includes: attribute selection, filling the missed values, correcting errors, or removing outliers (unusual or exceptional values), resolve data conflicts using domain knowledge or expert decision to settle inconsistency.

3.2.1 Attribute Selection

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 (Shearer, 2000). Therefore, in this thesis the attribute are selected with the help of domain expert and extensive literature review. Because taking all the variables in the data base we have, feed them to the data mining tool and find those which are the best predictors may be does not work very well. One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models (Two Crows Corporations, 2005). 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.

The national survey data set obtained contains many attributes and to decide on the relevant attributes for this study the researcher has discussed with domain expert in the area. As described in table 3.1, the following attributes are selected from the five years survey: Mother's age at birth, Birth order, Residence, Region, Mother's education, Wealth quintile, Size of child at birth, marital status, and Tetanus injections before Birth, Antenatal visits during pregnancy, Smoking status, and Anemia level, Sex of child, Abortion and others. The final selected attributes were prepared and preprocessed as stated in the following section, before developing the models.

Table 3.1: Description of the attributes selected from EDHS 2005 survey

No	Attributes	Description	Values	Data Type	Missing and unknown
1	Mother's Age	Age of mother at birth of the child	Mothers age in 5-year intervals(15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49)	Nominal	0
2	Birth order	Numbers of pregnancy including this birth	Numeric values of pregnancy including this one	Numeric	0
3	Residence	Type of place of residence.	Urban or Rural	Nominal	0
4	Region	The 11 administrative region of the country	Tigray, Afar, Amhara, Oromiya, Somali, Benishangul Gumz, Southern National Nationality People (SNNP), Gambela, Harari, Addis Abeba, Dire dawa	Nominal	0
5	Mother's education	The levels of mothers education during the birth	no education, primary, secondary, tertiary (Higher)	Nominal	0
6	Wealth index	The wealth of the mother during birth	Poorest, Poorer, Middle, Richer, Richest	Nominal	0
7	Marital Status	Marital status during pregnancy	Never married, Married, Living together, Widowed, Divorced, Not living together	Nominal	0
8	Religions	Religion of the mothers	Orthodox, Muslim, Protestant, Catholic, Tradition , other	Nominal	0
9	Sex	Sex of the child	Male or Female	Nominal	0
10	Ever had a terminated pregnancy	Pregnancies terminated before the last birth due to miss courage, abortion, or stillbirth.	No or yes	Nominal	2 (.02%)
11	Tetanus injections before Birth	women who received two or more tetanus injections during the pregnancy	It contains nominal values from 0 to 7 of tetanus injection before birth	Nominal	1227(12.4 %)
12	Antenatal visits during pregnancy	Number or Timing of Antenatal Visits during pregnancy	The numeric values of visit	Numeric	1450(14.7 %)
13	Smoking status	Smoking of cigarette during pregnancy	yes or no	Nominal	0
14	Birth Weight	The weight of the child after birth (this is the class attribute)	Normal or Low	Nominal	58(0.59%)
15	Iodine intake	The amount of iodine in the diet	0 ppm (no iodine), 7 ppm, 15ppm, and 30ppm	Nominal	228 (2.3%)

3.2.2 Statistical Summary of the Attributes (features)

Here the selected attributes used for model building are statistically described in details. This statistical summary of the attributes is helpful for understanding of the data set for experimentation.

Mother's Age

Mother's age is important demographic variables and is the primary basis of demographic classification in vital statistics, censuses, and surveys. It is also very important variables in the study of mortality and fertility. The age of mothers is classified by five year age groups. This attribute is categorized into seven parts as shown below table 3.2.

Table 3.2: Summary of Mother's Age Attribute

Mother Age : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
15-19	533	5.41
20-24	2062	20.91
25-29	2845	28.85
30-34	1992	20.20
35-39	1499	15.20
40-44	644	6.53
45-49	286	2.90
Missing	0	0.00
Total	9861	100

Place of Residence

Place of residence is nominal attributes; the possible values of this attribute are rural and urban. The nominal value of this attribute is described in table 3.3. As we can see the (Modal value) majority of respondents reside in rural areas.

Table 3.3: Statistical Summary of residence Attribute

Residence : Nominal		
Distinct Values	Frequency	Percent (%)
Urban	1358	13.8
Rural	8503	86.2
Missing	0	0.00
Total	9861	100

Religion

This nominal attribute has six distinct values (orthodox, catholic, protestant, Muslim, tradition and others. The detail summary of this attribute is described on table 3.4.

Table 3.4: Statistical Summary of Religion Attribute

Religion : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Orthodox	3901	39.6
Catholic	92	0.9
Protestan	1776	18.0
Muslim	3847	39.0
Tradition	151	1.5
Others	94	1.0
Total	9861	100.0

Mother's education

Mother’s education is indirectly related to a child’s health. Mothers Education is nominal attribute that contains four distinct values (No Education, Primary, Secondary, and higher). The most frequent value for Educational Level of the women is No Education as shown in table 3.5.

Table 3.5: Statistical summary of levels of mother Education Attribute

Mother’s education : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
No Education	7609	77.2
Primary	1548	15.7
Secondary	633	6.4
Higher	71	0.7
Missing	0	0.00
Total	9861	100

Marital Status

Table 3.6 shows the distribution of women by marital status. The marital status attributes is nominal. This attribute contains six distinct values married refers to both legal or formal marriage, living together refers to informal unions in which a man and a woman live together, never married, widowed, divorced , not living together. We can see from table 3.6 that the most frequent value for this attribute is married.

Table 3.6: Statistical Summary of Marital Status Attribute

Marital Status: Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Never married	49	0.5
Married	9078	92.1
Living together	171	1.7
Widowed	142	1.4
Divorced	267	2.7
Not living together	154	1.6
Missing	0	0.00
Total	9861	100

Region

The region attribute contains a total of 11 administrative region of the country. This attribute is nominal. The distinct values of region attribute are Tigray, Afar, Amhara, Oromiya, Somali, Benishangul Gumz, Southern National Nationality People (SNNP), Gambela, Harari, Addis Abeba and Dire dawa. The table 3.7 below shows the distribution of mothers by region.

Table 3.7: Statistical Summary of region Attribute

Region: Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Tigray	980	9.9
Afar	574	5.8
Amhara	1458	14.8
Oromiya	1938	19.7
Somali	663	6.7
Ben-Gumz	698	7.1
SNNP	1730	17.5
Gambela	515	5.2
Harari	514	5.2
Addis Abeba	380	3.9
Dire dawa	411	4.2
Missing	0	0.00
Total	9861	100

Iodine Test

Insufficient iodine in the diet can lead to serious health problems. Disorders arising from iodine deficiency range from goiter to mental and neurological disorders. Deficiency of iodine also causes abortion, stillbirth, low birth weight in infants, and premature birth. The principal cause of iodine deficiency is inadequate iodine in foods. Since iodine cannot be stored for long periods by the body, tiny amounts are needed regularly (100-150 micrograms per day per person). Salt that contains at least 15 parts per million (ppm) of iodine is considered to be adequately iodized. Iodine intake attribute is the nominal one that have four valid values. Table 3.8 shows the statistical distribution of iodine in salts. The missing values are replaced with modal values no iodine.

Table 3.8: Statistical Summary of Iodine intake Attribute

Iodine intake : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
0 parts per million (no iodine)	4256	43.2
7 parts per million	3290	33.4
15 parts per million	1627	16.5
30 parts per million	460	4.7
Missing	228	2.3
Total	9861	100

Smoking status

Mothers smoking status during pregnancy, the smoking status attribute is nominal type. The summary of this attribute is described in table 3.9 below.

Table 3.9: Statistical Summary of Smoking status Attribute

Smoking status : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Yes	325	3.3
No	9536	96.7
Missing	0	0
Total	9861	100

Abortions

This attribute refers to the previous termination of a pregnancy before birth, resulting in some reasons. Some abortions occur naturally because a fetus does not develop normally or because the mother has an injury or disorder that prevents her from carrying the pregnancy to term. This type of spontaneous abortion is commonly known as a miscarriage. Other abortions are induced that is, intentionally brought on because a pregnancy is unwanted or presents a risk to a woman's health, or because the fetus is likely to have severe physical or mental health problems.

The abortion attribute is nominal type that contains two distinct values (Yes and No). The statistical summary of this attribute is described as follows, the modal values of the attribute is 'no' then the missing values is replaced by 'no'

Table 3.10: Statistical Summary of Abortion Attribute

Abortion : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Yes	711	92.77
No	9148	7.21
Missing	2	0.02
Total	9861	100

Sex of Child

As we can see in table 3.11, sex of child is nominal attribute with possible values of male and female. The statistical distribution of this attribute is even as shown in table 3.11.

Table 3.11: Statistical Summary of sex of child Attribute

Sex of child : Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Male	5027	51
Female	4834	49
Missing	0	0
Total	9861	100

Birth order numbers

Birth order number is numeric value attribute that refers to the total number of births to this mother, including this birth. The frequent value of this attribute is one.

Table 3.12: Statistical Summary of Birth order numbers Attribute

Birth order Number: Numeric	
Minimum	1
Maximum	16
Mean	3.99
Standard Deviation	2.594
Mode	1
Missing	0

Tetanus Injections before Birth

Tetanus is an infectious disease caused by contamination of wounds from bacteria that live in the soil. The Tetanus Toxoid (TT) vaccine is given during your pregnancy to prevent tetanus to a mother as well as the baby. Antibodies formed in her body, after the vaccination, are passed on to her baby and protect her for a few months after birth. It also helps prevent premature delivery. A tetanus injection before birth attributes is nominal in type which contains 8 distinct values as shown in table 3.13. The modal value of this attribute is 0. Therefore the missing value is replaced by 0(modal value).

Table 3.13: Statistical Summary of Tetanus Injections before Birth Attribute

Tetanus Injections before birth: Nominal		
Distinct Values	**Frequency**	**Percent (%)**
0	6214	42.3
1	547	5.5
2	828	8.4
3	760	7.7
4	162	1.6
5	98	1.0
6	17	0.2
7	8	0.1
Missing	1227	12.4%
Total	9861	100.0

Antenatal visits during pregnancy

Antenatal visits during pregnancy are the clinical assessment of mother and fetus during pregnancy, for the purpose of obtaining the best possible outcome for the mother and child. This attribute is numeric and contains the numbers of mother's visits during pregnancy. The missing values of the attribute are 1450(14.7%), with the help of domain expert the missing value is replaced by mean values.

Table 3.14: Statistical Summary of Antenatal visits during pregnancy Attribute

Antenatal visits during pregnancy : Numeric	
Minimum	0
Maximum	20
Mean	1.361
Standard Deviation	2.462
Mode	0
Missing	1450(14.7%)

Birth Weight

This attribute is a class attribute that is transformed into nominal by considering weight of babies greater than 2500g as normal and below 2500g as low. The statistical summary of the attribute is described in table 3.15. As we can see the modal value is Normal, and the missing values are replaced by this value.

Table 3.15: Statistical Summary of Birth weight Attribute

Birth Weight: Nominal		
Distinct Values	**Frequency**	**Percent (%)**
Normal	7080	71.8
Low	2723	27.6
Missing	58	0.6
Total	9861	100.0

3.2.3 Handling Missing Values

For many real-world applications of data mining, even when there are huge amounts of data, the subset of cases with complete data may be relatively small. A number of problems are faced while bringing the data into proper format. Missing data is the most common problem that comes up during the data analysis process. Missing values

minimizing the accuracy of classification and rules generated by the selected data mining algorithm. Missing values lead to the difficulty of extracting useful information from that data set. Solving the problem of missing data is of a high priority in the field of data mining and knowledge discovery. Handling missing values by appropriate methods does not affect the quality of the data. In this thesis the two widely used methods are applied. One is avoid the missing data and other is data Imputation (Kantardzic, 2003).

Avoid the missing data is not time consuming and same time it is very easy to follow. But there are many drawback associated with this method. Deleting records may result in losing some information. If the sample data size is large avoiding some records or attributes may not affect the results, but still we need to keep in mind we are losing something.

Data imputation is another method of handling missing values. By using this method we try to fill missing values in the records and attributes. This method is quite useful because by following this method we can make sure we have all the information from responders. There are different approaches suggested for Data imputation in handling attributes with missing values in the data set. One approach among them is the use of attributes mean to fill in the missing values when the attribute type is numerical (Han and Kamber, 2006, Kantardzic, 2003). For example, the researcher has used the mean value (1.4) for the attribute Antenatal visits during pregnancy, which has about 14.7% of its data missing.

To handle the problem of missing values for categorical variables the suggestion is to group such fields into nominal and ordinal variables and to take the median for ordinal variables and to take the modal value for nominal variables (Two Crows Corporation, 2005). For example, for the variable Iodine intake which has about 2.3% missing, based on the above suggestion the modal value ‘0’ (0 ppm i.e no iodine) is substituted for the missing. Other variables which have missing data values have been treated in similar approach. The detail description of handling missing values data is shown in table 3.16.

Table 3.16: Handling of Missing Values

No	Attributes	Percentage of missing values	Replace with	Data Type	Remark
1	Ever had a terminated pregnancy	0.02%	0	Nominal	Mode
2	Tetanus injections before Birth	12.4%	0	Nominal	Mode
3	Antenatal visits during pregnancy	14.7%	1.4	Numeric	Mean
14	Birth Weight	0.6%	0(normal)	Nominal	Mode
15	Iodine intake	2.3%	0(0=ppm)	Nominal	Mode

3.2.4 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 (Chackrabarti, et al, 2009).

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) (Chackrabarti, et al, 2009).

In this research “mother age” attributes was discretized to reduce the unlike values of the attributes in order to obtain Knowledge (patterns), and to make the data set suitable for

mining tools. Then this attribute is later discretized in to seven bins, which is five- year's age group. Table 3.17 shows the discretized labels of mother's age attribute.

Table 3.17: Discretized Result of Mother Age attributes

Values (age groups)	**Frequency**
15-19	533
20-24	2062
25-29	2845
30-34	1992
35-39	1499
40-44	644
45-49	286
Total	**9861**

From the original data set the distinct value of mother age were 35. As we can see from the table 3.17 above, the attribute is reduced into seven labels through discretization using equal width (i.e five year intervals). Making seven distinct values would be easy to interpret the model.

Another attribute that need discretization is iodine intake attribute. The attribute have different values in original dataset but with the help of domain expert this attributes was reduced as follows.

Table 3.18: Iodine intake attribute from original data set

Labels	**frequency**
No iodine (0 parts per million)	4253
7 parts per million	3290
15 parts per million	1627
30 parts per million	460
No salt	3
Missing values	228
Total	9861

As we can see in table 3.18, originally iodine intake attributes has five distinct values. But with the help of domain expert the attribute is reduced into the first four distinct values. Therefore the distinct value of these attribute are 0 Parts per Million (no iodine), 7 parts per million, 15 parts per million, and 30 parts per million.

3.2.5 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 preprocessing 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 preprocessed data some basic statistics summary will be produced for each attributes. For categorical attributes, the frequency for each attribute value is shown. Moreover; the data that was converted into ARFF format is passed through important steps of data preprocessing mentioned in the previous section.

3.2.6 Setting the class attribute

In decision tree classification technique, which is supervised learning, predefined classes are required in order to train and build classification models. The setting of predefined class is done intentionally because the employed technique for this study is decision tree classification. In order to classify records into different classes the target attribute selected in this research is birth weight. It has got two different values – low birth weight and normal birth weight. Therefore the attribute is the dependent attribute while the rest of the variables are the independent attributes for this particular study.

3.2.7 Data type conversion

Before proceeding into building the models, some attributes and class attributes that were in numeric type were converted into nominal using Weka's attribute type converter in order to enable Weka's implementation of decision tree classifier and rule induction algorithm.

3.3 Model Building

According to the selected methodology of CRISP_DM the next step following data preparation is model building. The major activities model building includes selection of modeling technique, generating test design, building model and model assessment.

3.3.1 Selection of modeling technique

The initial step in model building is selection of specific modeling technique. The selection of modeling technique is based on the objective formulated. Since the purpose of this research is to develop a predictive model for low birth weight babies, classification algorithms has been used for building the model. The analyses were performed using WEKA environment. Inside the Weka system, there exist many classification algorithms which can be classified into two types; rule induction and decision-tree algorithms.

Rule induction algorithms generate a model as a set of rules. The rules are in the standard form of IF-THEN rules. Meanwhile, decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class outputs (Witten and Frank, 2005). Decision trees are one of the most widely used and practical forms of machine learning and data mining. Decision tree models are built by a process that is known as recursive partitioning. The classification algorithms used in this study were J48 and PART.

The J48 algorithm is the Weka implementation of the C4.5 top-down decision tree learner proposed by J. Ross Quinlan. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. The C4.5 decision tree algorithm is used in this research. This algorithm is implemented altering parameters such as confidence factor ,pruning and unpruning, changing the generalized and binary split decision tree classification options as shown in figure. 3.1. It is important to understand the variety of options available when using this algorithm, as it can make a significant difference in the quality of results. In many cases, the default settings will prove adequate, but in some cases, each choice may require some consideration (Witten and Frank 2005; Han and Kamber, 2006).

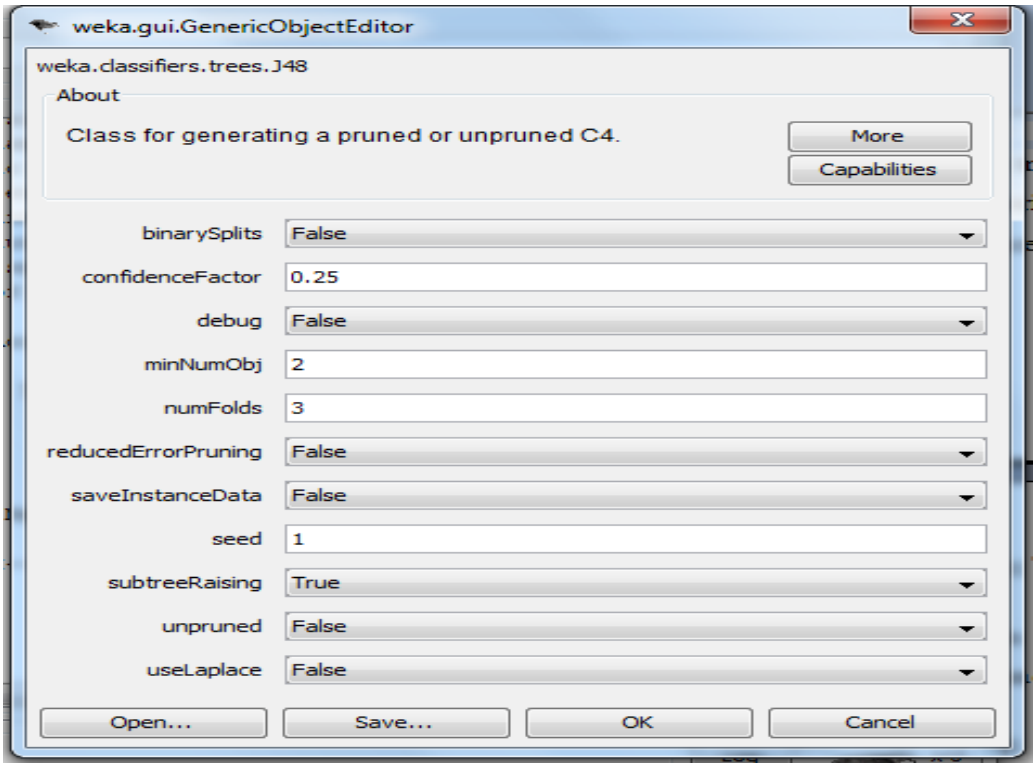

Figure 3.1: J48 Classifier Parameters Window in Weka software

The J48 algorithm gives several options as shown in figure 3.1 in related to tree pruning. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential over-fitting. The basic algorithm described above recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible. This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular habit of that data. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data. This is because pruning employs various means to relax the specificity of the decision tree, hopefully improving its performance on test data. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy (Witten and Frank 2005).

J48 employs two pruning methods. The first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf -- basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in

J48 is termed subtreeraising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Subtreeraising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtreeraising can be somewhat computationally complex (Witten and Frank 2005).

Error rates are used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over-fitting. This approach is known as reduced error pruning. Though the method is straight-forward, it also reduces the overall amount of data available for training the model. For particularly small datasets, it may be advisable to avoid using reduced error pruning (Witten and Frank 2005).

Other error rate methods statistically analyze the training data and estimate the amount of error inherent in it. The mathematics is somewhat complex, but this approach seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires a confidence threshold or ConfidenceFactor, which by default is set to 25 percent. This option is important for determining how specific or general the model should be. If the training data is expected to conform fairly closely to the data you'd like to test the model on, this figure can be lowered. The reverse is true if the model performs poorly on new data; try decreasing the rate in order to produce a more pruned (i.e, more generalized) tree (Witten and Frank 2005).

There are several other options that determine the specificity of the model. The minimum number of instances per leaf is one powerful option (MinNumObj). This allows us to dictate the lowest number of instances that can constitute a leaf. The higher the number, the more general the tree, lowering the number will produce more specific trees, as the leaves become more granular. The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using an

inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees. There is also an option available for using Laplace smoothing for predicted probabilities. Laplace smoothing is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities (Witten and Frank 2005).

The most basic parameter is the tree pruning option. Depending on how the training and test data have been defined that the performance of an unpruned tree may superficially appear better than a pruned one. As described above, this can be a result of over fitting. It is important to experiment with models by intelligently adjusting these parameters. Often, only repeated experiments and familiarity with the data will tease out the best set of options (Witten and Frank 2005).The five basic J48 parameters are described in the following table 3.19. This table contains name, default value, possible values and description of the parameters.

Table 3.19: Parameters for building J48 trees.

Name	Possible values	Default values	Description
M- number of instances	1,2,…	2	Minimum number of instances in leaves (higher values result in smaller trees)
U – unpruned trees	False/True	False	Use unpruned tree (the default value 'no' means that the tree is pruned)
C - confidence factor	10^{-7} - 0.5	0.25	Confidence factor used in postpruning (smaller values gain more pruning).
S – subtree raising	True/False	True	Whether to consider the subtree raising operation in postpruning.
B – use binary splits	False/ True	False	Whether to use binary splits on nominal attributes when building the tree.

PART is a separate-and-conquer rule learner proposed by (Witten and Frank 2005). The algorithm generates sets of rules called 'decision lists' which are ordered set of rules. PART builds a partial C4.5 decision tree in each iteration and converts the "best" leaf into a rule. The parameters mentioned in J48 decision tree algorithm are also applicable here.

3.3.2 Generation of test design

Numerous measures are used for rule evaluation in machine learning and knowledge discovery. In classification rule induction, the most frequently used measure is classification accuracy. Other standard measures include precision and recall, sensitivity and specificity.

Prior to building a model, a procedure needs to be defined to test the model's quality and validity. In supervised data mining tasks like classification, it is common to use error rates as quality measures for data mining models. Therefore, the test design specifies that the dataset should be separated into training and test set, the model is built on the training set and its quality estimated on the test set. The process of building predictive models requires a well defined training and validation protocol in order to ensure the most accurate and robust prediction (Two crows corporations, 2005). Weka mining software has the facility to extract a random sample and then test the accuracy of the classifier on disjoint collection of cases. For this study 10 -fold cross validation has been used. To evaluate the robustness of the classifier, the normal methodology is to perform cross validation on the classifier. 10 fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier (Witten and Frank, 2005).

CHAPTER FOUR

EXPERIMENTATION AND RESULT ANALYSIS

4.1 Model building using J48 Algorithms

In the CRISP_DM methodology, Model building is an iterative process. Therefore, it is important to conduct different experiments to find the best model for solving the problem. In this study, different experiments are conducted altering parameters of the J48 decision tree and PART rule induction algorithm for building the best predictive model. The parameter detail is discussed in the previous chapter. The experiment was done in two categories. In the first category, all 14 attributes were employed. In this category - eleven (11) possible experiments were conducted. Details setup of these experiments are shown in Table 4.2

Before selecting the values of parameters as shown in Table 4.2, the researcher conducted different experiment by altering the confidence factor values. Conducting different experiment by altering confidence factor value is important to produced optimal and accurate decision tree with minimized incorrectly classified instances. Table 4.1 shows the numbers of incorrectly classified instance for each confidence factor value.

Table 4.1: Values of incorrectly classified instance with each confidence factor values.

Confidence factor	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Incorrectly classified instance	2147	980	795	709	647	602	565	554	544	534	523

As can be seen in table 4.1 the increment of confidence factor is very important to obtain minimized incorrectly classified instance. Table 4.2 presented some selected confidence factors with other important J48 decision tree classifier parameters.

Table 4.2: Values of parameters used in the eleven Experiments

Experiments	Parameters		
	Pruned	**Confidence Factor**	**Numbers of Instance (minNumObj)**
Experiment #1	True	0.25	2
Experiment #2	True	0.25	5
Experiment #3	True	0.25	10
Experiment #4	True	0.30	2
Experiment #5	True	0.30	5
Experiment #6	True	0.30	10
Experiment #7	True	0.50	2
Experiment #8	True	0.50	5
Experiment #9	True	0.50	10
Experiment #10	False	0.25	2
Experiment #11	False	0.30	2

Each experiment is conducted using *k-fold* cross validation as it is appropriate whether the size of the data set is sufficiently large or not. The most common value of *k* recommended in different studies is 10. This is because extensive tests on numerous datasets, with different learning techniques, have shown that '10' is about the right number of folds to get the best estimate of error. In ten-fold cross validation, the training set is equally divided into 10 different subsets. Nine out of ten of the training subsets are used to train the learner and the tenth subset is used as the test set. The procedure is repeated ten times, with a different subset being used as the test set (Whitten and Frank, 2005).

There are imbalanced classes in this dataset and in order to solve this problem re-sampling technique is used for all experiments. Re-sampling is important to converts the imbalanced class distribution towards a uniform class distribution. The class imbalance problem has been known to hinder the learning performance of classification algorithms. A data set is called imbalanced if at least one of the classes is represented by significantly less number of instances than the others. In imbalanced data classification, the class boundary learned by the standard machine learning algorithms can be severely skewed toward the positive class. Thus, the false-negative rate can be excessively high. One major research direction to overcome the class imbalance problem is to resample the

original training data set. The underlying motivation for re-sampling methods is to provide the learner with a training set having more balanced classes (Nitesh et al, 2002)

As we can see from table 4.2 the selected parameters for the first experiment are confidence factor (0.25, 0.30, and 0.50), minNumObj (2, 5, and 10) and Unpruned (true/false). These three parameters are useful for optimizing accuracy and size of decision tree (pruning tactics) in testing set of data for model building.

After all necessary J48 decision tree parameters were set; the experiment listed in table 4.2 was conducted. The summery of obtained outputs of the eleven (11) different experiments is presented in Table 4.3 below.

Table 4.3: Performance report of the J48 Decision tree classifier

Performance Measure	**Experiments**										
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**
Accuracy (%)	93.9	89.6	**83.9**	94.3	89.9	84.2	**94.7**	90.3	84.5	92.8	92.8
Mean absolute Error	0.08	0.14	**0.22**	0.07	0.13	0.21	**0.06**	0.12	0.21	0.08	0.08
Numbers of leaves	1509	1093	**675**	1576	1174	685	**1673**	1228	776	2601	2601
Size of tree	1932	1381	**849**	2016	1480	863	**2133**	1542	971	3312	3312
Time taken to build(sec)	0.65	0.61	**0.47**	0.64	0.54	0.46	**0.47**	0.52	0.46	0.47	0.47
AV. TP Rate	0.94	0.90	**0.84**	0.94	0.90	0.84	**0.95**	0.90	0.85	0.93	0.93
AV. FP Rate	0.09	0.16	**0.26**	0.08	0.15	0.25	**0.08**	0.15	0.24	0.11	0.11
AV. Precision	0.94	0.89	**0.84**	0.94	0.90	0.84	**0.95**	0.90	0.84	0.93	0.93
AV. ROC Area	0.96	0.94	**0.83**	0.94	0.94	0.89	**0.97**	0.95	0.90	0.95	0.95
AV. Recall	0.94	0.90	**0.84**	0.94	0.90	0.84	**0.95**	0.90	0.86	0.93	0.93
CCI	9259	8831	**8269**	9296	8864	8304	**9338**	8902	8335	9155	9155
ICCI	602	1030	**1592**	565	997	1557	**523**	959	1526	706	706

4.2 Model Evaluation

When evaluating a classifier, there are different ways of measuring its performance. The experiments conducted above have been analyzed and evaluated in terms of evaluating classifiers performance values, accuracy, confusion matrix values, TP and FP Rate, number of leaves, and size of tree generated, ROC curves and execution time.

As shown in Table 4.3, performance of the classifier on the testing set increased as the confidence factor increased up to about 0.5 at a peak of 94.7% accuracy. Correctly and incorrectly classified instance at this accuracy are 9338 and 523 respectively from 9861 instance. From eleven different trials experiment #7 is the best model in terms of accuracy and minimized incorrectly classified instance. The rest of the performance evaluating classifiers is discussed as follows.

The Confusion Matrix of Experiment #7 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.

Table 4.4: Confusion Matrix

=== Confusion Matrix ===		
A	b	<-- classified as
6588	238	a = Normal
285	2750	b = Low

The above confusion matrix shows that 6588 instances were correctly predicted as normal birth weights (True positive). True positive of the actual class of the test instance is Normal weight and the classifier correctly predicts the class as Normal weight. The numbers of instance which were correctly predicted as low weights are 2750 instances (True negative). In this case of true negative the actual class of the test instance is low weight and the classifier correctly predicts the class as low weight. Therefore, correctly classified instance are the sum of this two figures (the sum of diagonal values of the table).

In contrast, table 4.4 shows that 285 instances were predicted as normal birth weight when they were in fact low birth weight (False Positives). False positive is when the actual class of the test instance is low weight but the classifier incorrectly predicts the class as normal weight. Lastly, the classifier predicted 238 instances as low weight (False Negatives). False negative is when the actual class of the test instance is Normal weight but the classifier incorrectly predicts the class as low weight. From the confusion matrix result it is possible to say the model was better at predicting low birth weight cases than the other experiments.

The following result has been extracted from Experiment #7 model. True Positive rate shows the percentage of low weight instances whose predicted values of the class attribute are identical with the actual values. FP rate shows the percentage of instances whose predicted values of the class attribute are not identical with the actual values.

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.965	0.094	0.959	0.965	0.962	0.969	Normal
	0.906	0.035	0.92	0.906	0.913	0.969	Low
Weighted Avg.	0.947	0.076	0.947	0.947	0.947	0.969	

If we take the first level where ‘weight=low’ TP Rate is the ratio of low weight cases predicted correctly to the total of positive cases, there were 2750 instances correctly predicted as low weight, and 3035 instances in all that were low weight. So the TP Rate (True Positive Rate) of low birth weight = 2750/3035 = 0.906. The FP Rate is then the ratio of normal weight of incorrectly predicted as low birth weight to the total of normal weight cases. 238 normal weight instances were predicted as low weight and there were 6826 normal weight in all. So the FP Rate is 238/6826 =0.035. We can follow the same method to calculate for ‘Weight=normal’ but as we can see from detailed accuracy by class TP Rate and FP Rate of Normal class level are 0.965 and 0.094 respectively. The model performance is good quality because it has high true positive rates with low false positive rates.

As can be seen from the detailed accuracy by class output, the ROC (Receiver Operating Characteristics) area of this model is highest (0.969). The Area under the ROC curve in figure 4.1 is higher. Higher numbers here indicates the model is the more accurate than the others. The ROC curve is a plot of how the classifier performs over the entire range of possible choices of cut-off values. Each point on the curve represents the True-Positive Rate plotted on the y-axis and the False-Positive Rate plotted on the x-axis that resulted from a particular cut-off value.

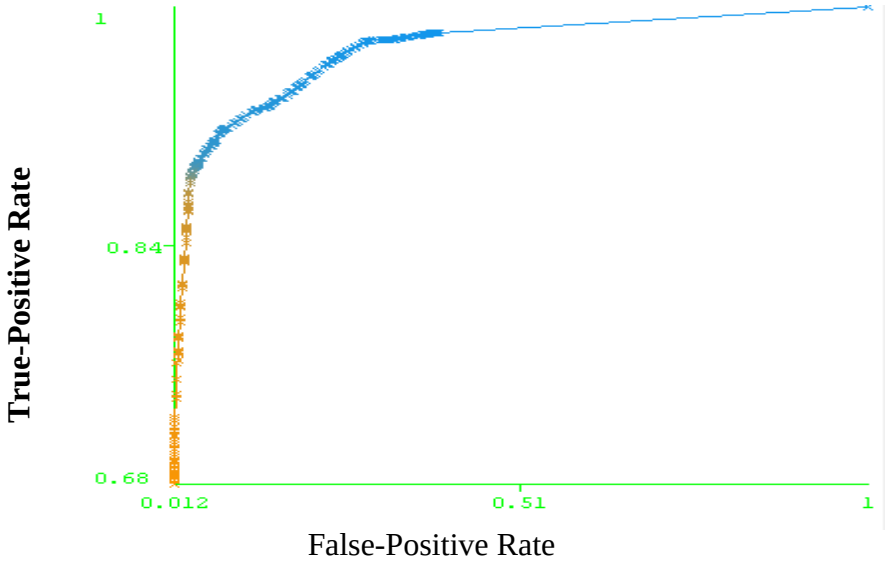

Figure 4.1: Experiment #7 ROC Area curve

As can be seen from the fig. 4.1 the larger the area under the ROC curve the more accurate the test. Proper utilization of pruning methods and techniques has shown to increase classification accuracy given an induced decision tree. But the size of the tree is very large and complex to interpret.

In the case of unpruned tree construction, the confidence factor has no effect for unpruned tree experiments. As we can see in table 4.3, Experiment #10 and #11, when confidence factor increase the values of the performance evaluators does not change. The values of model accuracy, True positive rate, ROC area curve, precision, and recall are less than the selected model (Experiment #7). But the numbers of leaves and size of tree is larger than the other experiments.

On the other hand, increasing the minimum number of instances per leaf (minNumObj) can reduce the tree size and numbers of leaves. In this study, the researcher found that minNumObj option is useful for reducing the size of the decision tree, but there is an impact on accuracy of the model. As can be seen on table 4.3 Experiment #3 is better than the other experiments with less numbers of leave and size of tree. Therefore, this parameter is very important for producing minimized size of trees for extracting rule easily.

The second category of the experiments is conducted using 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 4.5: List of attributes with their information gain

No.	Ranked attributes:	Information Gain	Deviation from maximum Info. Gain (Vmax - Vi)
1	5 Region	0.01157496	
2	13Antenatalvisitsduringpregnancy	0.0106206	0.000954
3	12 Tetanusinjectionsbeforebirth	0.00891656	0.002658
4	9 SexOfChild	0.00787056	0.003704
5	2 Birthorder	0.0077618	0.003813
6	6 MothersEducation	0.00727862	0.004296
7	4 Religion	0.00604402	0.005531
8	7 Wealthindex	0.00392049	0.007654
9	14 IodinTest	0.00326295	0.008312
10	10 Smokingstatus	0.0027464	0.008829
11	8 MaritalStatus	0.00224063	0.009334
12	3 Residence	0.00190286	0.009672
13	1 Age	0.0016666	0.009908
14	11 Abortion	0.00000948	0.011565

As shown in the Table 4.5, the selected attributes using entropy based information gain method of Weka are the first 10. Therefore, the last four attributes are reduced. The second round experiments is used these ten attributes. The result of the second round experiments is compared with the first round experiments and the best one is selected. Summary of the parameters set for each of the second round four experiments are presented in Table 4.6 below.

Table 4.6: Values of parameters used in the nine Experiments

Experiments	**Parameters**		
	Pruned	**Confidence Factor**	**Numbers of Instance (minNumObj)**
Experiment #1	True	0.25	2
Experiment #2	True	0.25	5
Experiment #3	True	0.25	10
Experiment #4	True	0.30	2
Experiment #5	True	0.30	5
Experiment #6	True	0.30	10
Experiment #7	True	0.50	2
Experiment #8	True	0.50	5
Experiment #9	True	0.50	10

As can be seen from table 4.6, the parameters are similar to the first round experiments. Therefore, in the second round nine experiments have been conducted. Hence the researcher performs the previous experiments using 10-folds cross validations, the testing option is also the same as the previous one. As mentioned before unpruned option is excluded. As can be seen in table 4.3, unpruned have no as such use for building the model. Summary of the outputs of the nine different experiments is presented in Table 4.7 below.

Table 4.7: Performance report of the J48 Decision tree classifier for reduced attributes

Performance Measure	Experiments								
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**
Accuracy (%)	92.5	88.4	**83.1**	92.9	88.7	83.4	**93.3**	88.9	83.6
Mean absolute Error	0.10	0.16	**0.23**	0.09	0.15	0.23	**0.09**	0.15	0.22
Numbers of leaves	1309	929	**566**	1350	968	607	**1384**	1026	668
Size of tree	1764	1235	**751**	1820	1284	803	**1866**	1351	876
Time taken (sec)	0.75	0.47	**0.50**	0.56	0.52	0.44	**0.55**	0.48	0.45
AV. TP Rate	0.93	0.88	**0.83**	0.93	0.89	0.83	**0.93**	0.89	0.84
AV. FP Rate	0.12	0.18	**0.27**	0.11	0.18	0.26	**0.10**	0.17	0.25
AV. Precision	0.92	0.88	**0.83**	0.93	0.89	0.83	**0.93**	0.89	0.83
AV. ROC Area	0.95	0.93	**0.87**	0.96	0.93	0.88	**0.96**	0.93	0.88
AV. Recall	0.93	0.88	**0.83**	0.93	0.89	0.83	**0.93**	0.89	0.84
CCI	9120	8718	**8190**	9163	8749	8226	**9197**	8765	8241
ICCI	741	1143	**1671**	698	1112	1635	**664**	1096	1620

As illustrated in the output summery of Table 4.3 and Table 4.7, the model has got the best performance during Experiment #7 with all attributes compared to other experiments with reduced and all attributes. However, it has been found that in all of the experiments conducted using the reduced attributes the performance of the model has declined than when the experiment was conducted using the whole attributes. On the other hand it is possible to observe that the reduction in the number of the attributes has also somehow reduced the complexity of the tree but this is not as much significant. The missing values imbalanced instance and the nature of the dataset may affect the accuracy of the models.

The second experiments with reduced attributes of details accuracy are smaller than the previous experiments. But these experiments have less numbers of leaves and size of tree than the previous experiments. Correctly and incorrectly classified instances are getting smaller with reduced attribute. The time taken in building the model is somewhat reduced. The other performance evaluators like ROC area, TP Rate, Precision, Recall are also less than the previous one. However, FP Rate is higher than the previous experiments.

To sum up, the performance comparison of 20 different experiments was also presented. As described in details in section 4.3 the overall best performance was achieved by J48

classifier, using pruned technique, by setting confidence factor (0.5), numbers of instance (2) and all 14 attributes data set. With a recall (true positive rate) of 95%, a false positive rate of 1%, a precision (positive predictive value) of 95%, and an accuracy of 94.7%. The analysis of the experimental results shows that the model is quite effective and efficient in detecting low birth weight.

Even though the achieved accuracy was very good, the attempt to increase the accuracy of the model further beyond this one is not successful and possible. One possible reason is that the data set is not uniformly distributed (i.e. there is imbalanced class). The second reason may be modal values used as a replacement for missed values might also be another factor for the prediction of some classes.

4.3 Generating Rules from J48 Decision Trees

In knowledge discovery, it is crucial to investigate interaction between attributes in order to induce interesting and useful prediction rules. Decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute. From the decision tree generated by the best model mentioned in section 4.3, it is possible to extract important rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node. The following are some of the most important rules/patterns extracted from the decision tree model.

- If Antenatal visits during pregnancy is “less than or equal to seven”, Smoking status the mother is “no”, sex of child is “Male”, their marital Status during pregnancy is “married”, their education status “no education” at all, Tetanus injections before birth is “zero”, their region is “Amhara”, the religion of the mother is “Orthodox” and Wealth index(economical status of the mother) is “middle” and the numbers of Birth order is “between one and four” then low birth weight baby will be born.
- If antenatal visits during pregnancy are “more than seven” times and their current marital status is “divorced” then low birth will occurred.

- If antenatal visits during pregnancy “less than or equal to seven”, Smoking status the mother is “no”, sex of child is “Male”, Marital Status during pregnancy is “Married”, their educational status “no education” at all, Tetanus injections before birth is “zero”, the region of the mother is “Somali” and contents of iodine in salt is “no” at all then low birth weight babies will expected.
- If antenatal visits during pregnancy are “less than or equal to seven”, Smoking status the mother is “no”, sex of child is “Male”, Marital Status during pregnancy is “Married”, the Mothers education “no education” at all, tetanus injections before birth is “one”, the religion of the mother is “orthodox” and numbers of Birth order is “less than or equal to four” then the low birth weight baby will born.
- If antenatal visits during pregnancy is ” less than or equal to seven”, smoking status the mother is “no , sex of child is “Male”, whose religion is “Muslim”, their residence is “rural”, current marital status is “married”, Tetanus injections before birth is “no” , mother region is “oromiya”, mothers age is between “25 and 29”, numbers of birth order is “less or equal to three” and the presence of iodine in salt is “7 part per million” then the probability of low weight baby is high.
- If antenatal visits during pregnancy are “less than or equal to seven”, smoking status the mother is “yes” , numbers of birth order is “more than five” and their age between “35 and 39” then mothers will born low weight baby.
- If antenatal visits during pregnancy is “less than or equal to seven”, Smoking status the mother is “no”, Sex of Child is “female”, mothers religion is “protestant” and their age is between “45 and 49”, then mothers will deliver low weight baby.
- If antenatal visits during pregnancy is between “0 and 7”, smoking status of the mother is “no”, sex of child is “female”, mothers religion is “orthodox”, presence of iodine in salt is “15 part per million” and numbers of birth order is “less or equal to three” then low weight baby will be expected.

- If antenatal visits during pregnancy “less than or equal to seven”, smoking status of the mother is “no”, sex of child is “female”, mothers religion is “orthodox”, Tetanus injections before birth is “zero”, the presence of iodine in salt is “no” and mothers age is between “20 and 24” then low weight baby will born.
- If antenatal visits during pregnancy is “less than or equal to seven”, smoking status of the mother is “no”, sex of child is “female”, mothers religion is “orthodox”, Tetanus injections before birth is “two”, mothers educational status “no eduction” at all and their age is between “15 and 19” then they will born low birth weight child.
- If antenatal visits during pregnancy is “less than or equal to seven”, smoking status the mother is “no”, sex of child is “female”, mothers religion is “Muslim”, their residence is “rural” and their current marital status is divorced then the probability of born low birth weight baby is high.

In general interesting rules were generated by J48 algorithm presented above. The pattern /rules obtained by setting pruned, confidence factor 0.25 and numbers of instance 10 J48 decision tree with all attributes model is useful for health care professionals to reduced low birth weight.

4.4 Model Building PART Rule Induction Algorithm

The second data mining technique used in this research was PART Rule induction algorithm. As mentioned in previous chapter, section 3.3.1, there are many classification algorithms which can be classified into two types; rule induction and decision-tree algorithms. Rule induction method apply an iterative process consisting of first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover. The final rule set is the collection of the rules discovered at every iteration of the process. The rules are in the standard form of IF-THEN rules.

The researcher prefers PART rule induction algorithms over other rule induction algorithms because it has the ability and potential to produce accurate and readable rules. PART is a separate-and-conquer rule learner proposed by (Witten and Frank 2005). The algorithm produces sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule. PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning.

The experiment is performed using all 15 attributes including class instance. The parameters of PART rule induction algorithms are default (i.e confidence factor =0.25, MinNumObj=2 unpruned = false). Running the PART rule induction algorithm on the supplied dataset generates rules in plain text form, which is simple to understand and interpret. The following texts show the summary of this experiment.

=== Run information ===

Number of Rules: 460

Time taken to build model: 9.9 seconds

=== Summary ===

Correctly Classified Instances	9304	94.3515 %
Incorrectly Classified Instances	557	5.6485 %
Kappa statistic	0.8663	
Mean absolute error	0.0737	
Root mean squared error	0.2261	
Relative absolute error	17.3066 %	
Root relative squared error	48.9915 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.966	0.107	0.953	0.966	0.959	0.961	Normal
	0.893	0.034	0.921	0.893	0.907	0.961	Low
Weighted Avg.	0.944	0.084	0.943	0.944	0.943	0.961	

=== Confusion Matrix ===

a	b	<-- classified as
6593	233	a = Normal
324	2711	b = Low

As can be seen from the above result, PART rule induction algorithm built model has an accuracy of 94.35%. Correctly and Incorrectly Classified Instances are 9304 and 557 respectively with recall (true positive rate) of 94%, a false positive rate of 8%, a precision (positive predictive value) of 94.3% and ROC curve area 96.1%.

As can be seen from above run information, the numbers of rules produced by PART algorithm are 460. These rules generated using PART algorithm are more clear and understandable. Some of the most important rules produced by PART rule induction algorithm are presented as follows.

- If Antenatal visits during pregnancy is “equal to one”, marital status of mothers is “divorced”, the place of residence is “rural”, birth order numbers is” less than or equal to three”, educational status of the mothers is “no education” at all, and the economical status of the mothers is “poorest” then 96.1% of the mothers will born low birth weight baby.

- If abortion is “no”, the presence of iodine in salt is “7 parts per million”, Antenatal visits during pregnancy is ‘less than or equal to one”, the economical status of the mothers is “poorer” and the region is “oromiya” then most likely low birth weight child will born.
- If the sex of child is “male”, numbers of birth order is ‘less than or equal to two”, the place of residence is “rural”, Tetanus injections before birth is “greater than or equal to one”, marital status of the mothers is “married” and Wealth index is “middle” then the probability of born normal weight baby is 92.1%.
- If the region the mothers is” Amhara”, wealth index of the mothers is “poorest” the presence of iodine in salt “not” at all and sex of child is “male” then low birth weight will happen.
- If abortion is “no”, marital status of the mother is “married”, the region of mother is “amhara”, wealth index of mother is “richer”, numbers of birth order is “less than or equal to three “and Tetanus injections before birth is “greater than one” then the birth weight of the child is normal.
- If the region the mothers are “Amhara” , the age of mothers is between “25 and 29”, educational status of the mothers is “no education” at all, wealth index is “middle”, the presences of iodine in salt is “no” and sex of child is “female” then the baby’s birth weight will low.
- If marital status is “married”, sex of child is “male”, educational status of mothers is “secondary”, tetanus injections before birth is “three”, Antenatal visits during pregnancy is “greater than two” then normal birth weight baby will born.
- If the region the mothers are “Amhara”, Antenatal visits during pregnancy is “equal to one” and educational status of the mothers is “no education” at all, numbers of birth order is “less than or equal to eight”, the economical status of the mothers is “poorest”, presences of iodine in salt is “no” and sex of child is “female” then the born child will have low weight.

- If marital status is “married”, educational status of the mothers is “secondary”, the region of the mothers is “Addis Ababa” then the weight of the child will be normal
- If marital status is “married”, mother age is between “45 and 49” and sex of child is “male” then low birth weight will happen.
- If marital status is “married”, the region of mother is “oromiya”, the religion of the mothers is “Muslim”, educational status of the mothers is “primary” and wealth index is “middle” then birth weight of child will be normal.

The comparison of best model built by J48 decision tree classifier and PART rule induction algorithms result is presented in table 4.8.

Table 4.8: Accuracy and number of rules /size produced by J48 and PART algorithms

Performance measures	J48 decision tree	PART Rule induction
Accuracy (in %)	94.7%	94.35%
Size of tree/ numbers of rules	2133	460

It can be seen from the summary of both algorithm results in table 4.8; J48 decision tree and PART rule induction algorithm have the most accuracy (having average of 94%). The researcher prefers J48 over PART, since the accuracy the model build using J48 decision tree is better than PART rule induction algorithm.

Analysis of both algorithms show that in this data set the attributes such as antenatal visits during pregnancy (antenatal care for pregnancy), mother’s educational level, and marital status, Iodine contents in salt, region, 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.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Summary

Low birth weight (LBW) is the main factor determining neonatal and prenatal survival and it is also associated with many adverse outcomes in newborn. Low birth weight has been defined by the World Health Organization (WHO) as weight at birth of less than 2,500 grams. This practical cut-off for international comparison is based on epidemiological observations that infants weighing less than 2,500 grams are approximately 20 times more likely to die than heavier babies. More common in developing than developed countries, a birth weight below 2,500 grams contributes to a range of poor health outcomes.

Low birth weight is one of the critical issues in Ethiopia that causes many babies short-term and long-term health consequences and tend to have higher mortality and morbidity. DHS Ethiopia report shows that the percentage of low birth weight babies has increased in the past five years from 8 percent in 2000 to 14 percent in 2005. The percentage of babies assessed by mothers as being very small at birth has increased over the same period from 6 percent to 21 percent

Low birth weight is a reasonable well-defined problem caused by factors that are potentially modifiable and the costs of preventing them are well within reach, even in under developing countries like Ethiopia. Therefore, it is very important to predict LBW in various communities in the country in order to come up with feasible intervention strategies to minimize the problem. This study is conducted to explore the potential applicability of data mining technology to predict low birth weight baby using EDHS 2005 (Ethiopia Demographic Health Survey) data set. A model was built using data mining technique to understand and address the factors associated with low birth weight.
Data was collected from Measure DHS Ethiopia. The methodology applied in this research was CCRISP-DM, which contains six major phases: business understanding, data understanding, and data preparation, model building, evaluation and deployment. A

total of 9861 records were used for the experiments. Some numeric values attributes are discretized using ten-bin discretization implemented in Weka. Besides, missing values are also processed using the mechanism in Weka, which replaces all missing values with the modes and means from the training instances.

The data mining techniques selected for predicting low birth weight was classification. J48 decision tree classifier and PART rule induction algorithms were selected for experiments. Because they are reliable and effective decision making techniques which provide high classification accuracy with a simple representation of gathered knowledge. When using these algorithms, the decision making process itself can be easily validated by an expert.

Several models were built implementing the J48 decision tree classifier algorithm of C4.5 and PART rule induction algorithms. These experiments are done using pruning with all and reduced attributes, by giving J48 classifiers parameters of different values. The researcher compared the classification performance of the decision trees with tree prunning and without tree prunning, and found that tree prunning can significantly improve decision tree's classification performance.

5.2 Conclusion

The performance comparison of those different experiments was also presented. Cross validation folds for the testing set (crossValidationFolds) was held at 10. The study found that increasing the minimum instance (minNumObj) requirement decreased the accuracy of the classifier and reduced the numbers of leaves and sizes of the tree. On the other hand, lowering the confidence in the training data (confidenceFactor) does not only reduce the tree size, but also helps in filtering out statistically irrelevant nodes that would otherwise lead to classification errors. It is possible to conclude that, several values for the confidence factor should be tested when generating decision trees to find the most appropriate value for the particular training set under examination.

The overall best performance was achieved by J48 decision tree classifiers using pruned technique , confidence factor of 0.5 , minimum numbers of instance (minNumObj) at 2

with all attributes data set, with a recall (true positive rate) of 95%, a false positive rate of 1%, a precision (positive predictive value) of 95%, and an accuracy of 94.7%. The analysis of the experimental results shows that the model is quite effective and efficient in detecting low birth weight.

The second algorithm conducted in this research is PART rule induction algorithms. The result show accuracy of 94.35% and correctly and Incorrectly Classified Instances are 9304 and 557 respectively and with recall (true positive rate) of 94%. A false positive rate of 8%, a precision (positive predictive value) of 94.3% and ROC curve area 96.1% . The numbers of rules produced by PART algorithm were 460. These rules are found to be more clear and understandable.

In general, the results from this study were interesting and encouraging; it can be used as decision support for health practitioner. The extracted rules in both algorithms are very effective for the prediction of low birth weight. From both algorithms, we can observe 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.

5.3 Recommendation

The results of this study have shown that data mining technology, particularly the J48 decision tree algorithm and PART rule induction algorithm are applicable for low birth weight prediction. The selected experiments conducted in this study were effective and efficient in detecting low birth weight and the extracted rules in both the algorithms are very effective for the prediction.

The researcher highly believes that the findings of this research can be used by health authorities to further explore the determinant of low birth weight in Ethiopia. The following recommendations are based on the result of this study and nature of the data set.

- Both experiments conducted using J48 decision tree and PART rule induction algorithms produced efficient models and interpretable rules (patterns). Hence it

is important for health authorities to utilize (deploy) the model developed with these data mining technique in order to use as a decision support tool in the identification of low birth weight determinant factor patterns.

- The selected algorithms and missing values might confuse the modeling process. However, the researcher believes that these variables might have potential to develop better accuracy models. Therefore further researches should be conducted using balanced, other algorithms and using more attributes to assess the applicability of data mining technique to predict low birth weight determinant patterns.
- Hospital and health centers especially those give delivery services for mothers must keep records properly that includes all mothers and child information, because these records are useful to predict the low birth weight better than survey data.
- Furthermore, it is very essential to assess the applicability of data mining techniques in predicting low birth weight determinant factors by using other data sets (such as Hospitals delivery catalog book records that contains biological risk factors in addition to demographic data set) which is more comprehensive to identifying the biological as well as demographical factors of low birth weight patterns.

References

Chakrabarti .S ,Earl C., Eibe F., Ralf H.G., Jaiwei H. , Xia J., Micheline K., Sam S. L.,Thomas P. ,Richard E. ,Dorian P., Mamdouh R.,Markus S.,Toby J. and Witten H. (2009). *Data mining know it all*. Morgan Kaufmann Publishers 30 Corporate Drive, Suite 400 Burlington, Unite State

Chapman P., Julian C., Randy K.,Thomas K., Thomas R.,Colin S., and Rüdiger W. (2000). CRISP-DM 1.0 - Step-by-step data mining guide. Available at www.spss.ch/file.php?file=/upload/1107356429_CrispDM1.0.pdf (Accessed on 23 February, 2011)

Colin Shearer. (2000). *The CRISP-DM Model: The New Blueprint for Data Mining.* Journal of Data Ware Housing Volume 5, Number 4.

Daniel T. Larose. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley & Sons, Hoboken, New Jersey.

David, Hand, Heikki Mannila, and Padhraic Smyth. (2001). *Principles of Data Mining.* The MIT Press, Cambridge, Massachusetts, London England.

David L. Olson , Dursun Delen.(2008). *Advanced Data Mining Techniques*. Springer-Verlag Berlin Heidelberg.

Dunham M. H. and Sridhar S. (2006). *Data Mining: Introductory and Advanced Topics*, Pearson Education, New Delhi.

Ethiopia Demographic and Health Survey 2005 report Central Statistical Agency Addis Ababa, Ethiopia, 2006.

Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996). *From Data Mining to Knowledge Discovery in Databases.* Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.42.1071&rep=rep1&type=pdf (Accessed on 5 January, 2011)

Fikre, E. and Aklilu M. (2000). *Change in birth-weight of Hospital-Delivered neonate in Addis Ababa.* Ethio. J. Health Dev, Addis Ababa.

Harleen Kaur and Siri Krishan Wasan. (2006). *Empirical Study on Applications of Data Mining Techniques in Healthcare*. Journal of Computer Science , New Delhi ,Volume 2: 194-200.

Ian H. Witten and Eibe Frank. (2005). *Data mining Text book: Practical Machine Learning tools and Techniques*. 2nd Ed. Morgan Kaufmann Publishers, San Francisco

Invnka O., Henrik j., Lone S. and Jorn O. (2007). *Maternal vaccination and preterm birth: using data mining as a screening too*l. Pharm World Sci, Volume 29:205–212

Jiawie, Han and Micheline Kamber. (2006). *Data mining Concept and Techniques.* 2nd Ed. Morgan Kaufmann Publishers, San Francisco

Koh H. and Tan G. (2006). *Data Mining Applications in Healthcare.* Journal of Healthcare Information Management, Volume. 19, Number 2: 64-72.

Kramer M.S. (1987). *Determinants of Low Birth Weight: Methodological assessment and meta-analysis,* Bulletin of the WHO, Volume 65, Number 5.

K.Srinivas , Kavihta.B and Govrdhan.A. (2010). *Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks*. (IJCSE) International Journal on Computer Science and Engineering, Volume 2, Number 2:194-200.

Laviniu Aurelian Badulescu (2007). *The choice of the best attributes selection measure in Decision Tree induction*. Annals of University of Craiova, Math. Comp. Sci. Ser. Volume 34, Number 1: 88-93.

Liangxiao J. and Chaoqun L. (2010). *An Empirical Study on Attribute Selection Measures in Decision Tree Learning*. Journal of Computational Information Systems volume 6, Number 1:105-112.

Lori Bowen Ayre. (2006). *Data Mining for Information Professionals*. Available at http://techessence.info/files/Ayre_DataMiningForInformationProfessionals_June2006.pdf (accessed on 3 February, 2011).

Max Bramer. (2007). *Principles of Data Mining: Undergraduate Topics in Computer Science*. Digital Professor of Information Technology University of Portsmouth, UK Springer-Verlag, London.

Mehmed Kantardzic (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* John Wiley & Sons, New Jersey.

Michael J.A Berry, and Gordon S. Linoff. (2004). Data *mining Techniques: For marketing, Sales, and Customer Relationship*. 2nd Ed., Wiley Publishing, Inc., Indianapolis, Indiana.

Michael Gutkin.(2008). *Feature selection methods for classification of gene expression profiles*. Tel-Aviv University, Faculty of Exact Sciences, School of Computer Science. Available at: http://acgt.cs.tau.ac.il/theses/Msc-Thesis-Michael-Gutkin.pdf

Mirjana P. B. and Dijana Ć. (2008). *Data mining usage in health care management: literature survey and decision tree application*. Medicinski Glasnik, Volume 5, Number 1:57-64.

Moheb A., Chris B., Dean A., Andrew J., Marianne J. and Olga P. (2005). *ezDataMiner and the Strategic Advantages of Data Mining.* Central Connecticut State University, USA.

Nitesh V., Kevin W., Lawrence O. Hall and W. Philip.(2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, Volume 16: 321-357.

Ping-Ning Tan, Michael Steinbach, Vipin Kumar. (2006). *Introduction to Data mining*. Pearson Educ. Inc.

Philip Baylis. (1999). *Better health care with data mining.* SPSS Inc. WPDMHC-0699.

Plate T., Band P., Joel B. and John G. (1997). *A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer*. Advances in Neural Information Processing, MIT Press.

Quinlan J. R. (1986). *Induction of decision trees*. Kluwer Academic Publishers, Boston - Manufactured , Netherlands.

Romero C., Sebastián V., Pedro G. and César H.(2008). Data Mining Algorithms to Classify Students. Computer Science Department. Córdoba University, Spain

Ruben D. and Canlas Jr. (2009). *Data Mining In Healthcare: Current Applications and Issues.* Carnegie Mellon University Australia. Available at: http://mines.humanoriented.com/classes/2010/fall/csci568/papers/Data_Mining_Health.pdf (Accessed on 5 January, 2011).

Santos M.F. and Azevedo C. (2005). *Data Mining*: Descoberta de Conhecimento em Bases de Dados, FCA Editora, Lisbon.

S. P. Deshpande and V. M. Thakare (2010). *Data Mining System and Applications: A Review*. International Journal of Distributed and Parallel systems (IJDPS) Volume 1, Number 1: 445-463

Tema T. (2006). Prevalence and Determinants of Low Birth Weight in Jimma Zone, Southwest Ethiopia. East African Medical Journal. Volume 83.

Thair Nu Phyu (2009). *Survey of Classification Techniques in Data Mining.* Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Volume 1.

Thearling K., Alex B.and Stephen S.(2010). *An Overview of Data Mining Techniques*. Available at: http://www.thearling.com/text/dmtechniques/dmtechniques.htm (Accessed on 10 February, 2011)

Two Crows Corporation. (2005). *Introduction to Data Mining and Knowledge Discovery. 3rd* Ed. Two Crows Corporation. 500 Falls Road, Potomac, USA

UNICEF/WHO. (2004). *Low Birth weight: Country, regional and global estimates*. UNICEF, New York.

Varun Kumar, Dharminder Kumar, and R.K. Singh. (2008). *Outlier Mining in Medical Databases: An Application of Data Mining in Health Care Management to Detect Abnormal Values Presented in Medical Databases.* International Journal of Computer Science and Network Security, Volume 8, Number 8.

V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman (2002). *Decision trees: an overview and their use in medicine*, Journal of Medical Systems, Kluwer Academic/Plenum Press, Volume 26, Number 5: 445-463.

Walter Alberto Aldana. (2000). *Data Mining Industry: Emerging Trends and New Opportunities.* Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, May 2000.

WHO Report. (1992). *International statistical classification of diseases and related health problems, tenth revision*, World Health Organization, Geneva.

WHO Technical Consultation (2006). *Promoting optimal fetal development.* Geneva, Switzerland .

WHO Technical Consultation. (2004). *towards the development of a strategy for promoting optimal fetal growth', Report of a meeting* (draft), World Health Organization, Geneva,

Xiaohua Hu (2003). *DB-HReduction: A Data Preprocessing Algorithm for Data Mining Applications.* College of Information Science and Technology, Drexel University Philadelphia, PA 19104, U.S.A.

Annexes

Annex A: the result of run information of first experiments using J48 with all attributes

Experiment #1: using J48 Algorithm *with all attributes*

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Birth Weight-

Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves : 1509

Size of the tree : 1932

Time taken to build model: 0.65 seconds

=== Summary ===

Correctly Classified Instances	9259	93.8951 %
Incorrectly Classified Instances	602	6.1049 %
Kappa statistic	0.8553	
Mean absolute error	0.0799	
Root mean squared error	0.2314	
Relative absolute error	18.745 %	
Root relative squared error	50.1367 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.964	0.117	0.949	0.964	0.956	0.964	Normal
	0.883	0.036	0.916	0.883	0.899	0.964	Low
Weighted Avg.	0.939	0.092	0.939	0.939	0.939	0.964	

=== Confusion Matrix ===

a	b	<-- classified as
6579	247	a = Normal
355	2680	b = Low

Experiment #2 using J48 Algorithm *with all attributes*

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.25 -M 5

Relation: Birth Weight-

Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves : 1093

Size of the tree : 1381

Time taken to build model: 0.61 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8831	89.5548 %
Incorrectly Classified Instances	1030	10.4452 %
Kappa statistic	0.7498	
Mean absolute error	0.1401	
Root mean squared error	0.2891	
Relative absolute error	32.8891 %	
Root relative squared error	62.6415 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.941	0.206	0.911	0.941	0.926	0.937	Normal
	0.794	0.059	0.856	0.794	0.824	0.937	Low
Weighted Avg.	0.896	0.161	0.894	0.896	0.894	0.937	

=== Confusion Matrix ===

a	b	<-- classified as
6421	405	a = Normal
625	2410	b = Low

Experiment #3 using J48 Algorithm *with all attributes*

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.25 -M 10
Relation: Birth Weight-
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 675
Size of the tree : 849

Time taken to build model: 0.47 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	8269	83.8556 %
Incorrectly Classified Instances	1592	16.1444 %
Kappa statistic	0.6038	
Mean absolute error	0.2211	
Root mean squared error	0.3484	
Relative absolute error	51.8793 %	
Root relative squared error	75.4733 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.918	0.341	0.858	0.918	0.887	0.879	Normal
	0.659	0.082	0.782	0.659	0.715	0.879	Low
Weighted Avg.	0.839	0.261	0.835	0.839	0.834	0.879	

=== Confusion Matrix ===

a	b	<-- classified as
6269	557	a = Normal
1035	2000	b = Low

Experiment #4 using J48 Algorithm *with all attributes*

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.3 -M 2
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
Number of Leaves : 1576
Size of the tree : 2016

Time taken to build model: 0.64 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	9296	94.2704 %
Incorrectly Classified Instances	565	5.7296 %
Kappa statistic	0.8645	
Mean absolute error	0.075	
Root mean squared error	0.2254	
Relative absolute error	17.6057 %	
Root relative squared error	48.8406 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.965	0.106	0.953	0.965	0.959	0.966	Normal
	0.894	0.035	0.918	0.894	0.906	0.966	Low
Weighted Avg.	0.943	0.085	0.942	0.943	0.942	0.966	

=== Confusion Matrix ===

a	b	<-- classified as
6584	242	a = Normal
323	2712	b = Low

Experiment #5 using J48 Algorithm *with all attributes*

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.3 -M 5
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 1174
Size of the tree : 1480

Time taken to build model: 0.54 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	8864	89.8895 %
Incorrectly Classified Instances	997	10.1105 %
Kappa statistic	0.7585	
Mean absolute error	0.1342	
Root mean squared error	0.284	
Relative absolute error	31.4908 %	
Root relative squared error	61.5262 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.941	0.196	0.915	0.941	0.928	0.941	Normal
	0.804	0.059	0.858	0.804	0.83	0.941	Low
Weighted Avg.	0.899	0.154	0.898	0.899	0.898	0.941	

=== Confusion Matrix ===

a	b	<-- classified as
6423	403	a = Normal
594	2441	b = Low

Experiment #6 using J48 Algorithm

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.3 -M 10
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 685
Size of the tree : 863

Time taken to build model: 0.46 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	8304	84.2105 %
Incorrectly Classified Instances	1557	15.7895 %
Kappa statistic	0.6145	
Mean absolute error	0.2146	
Root mean squared error	0.3436	
Relative absolute error	50.3653 %	
Root relative squared error	74.4353 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.917	0.326	0.863	0.917	0.889	0.885	Normal
	0.674	0.083	0.783	0.674	0.724	0.885	Low
Weighted Avg.	0.842	0.251	0.839	0.842	0.839	0.885	

=== Confusion Matrix ===

a	b	<-- classified as
6259	567	a = Normal
990	2045	b = Low

Experiment #7 using J48 Algorithm

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.5 -M 2
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 1673
Size of the tree : 2133

Time taken to build model: 0.47 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	9338	94.6963 %
Incorrectly Classified Instances	523	5.3037 %
Kappa statistic	0.875	
Mean absolute error	0.0672	
Root mean squared error	0.2171	
Relative absolute error	15.7706 %	
Root relative squared error	47.0358 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.965	0.094	0.959	0.965	0.962	0.969	Normal
	0.906	0.035	0.92	0.906	0.913	0.969	Low
Weighted Avg.	0.947	0.076	0.947	0.947	0.947	0.969	

=== Confusion Matrix ===

a	b	<-- classified as
6588	238	a = Normal
285	2750	b = Low

Experiment #8 using J48 Algorithm

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.5 -M 5
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 1228

Size of the tree : 1542

Time taken to build model: 0.52 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	8902	90.2748 %
Incorrectly Classified Instances	959	9.7252 %
Kappa statistic	0.7682	
Mean absolute error	0.1283	
Root mean squared error	0.2782	
Relative absolute error	30.1085 %	
Root relative squared error	60.2656 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.942	0.186	0.919	0.942	0.931	0.945	Normal
	0.814	0.058	0.862	0.814	0.837	0.945	Low
Weighted Avg.	0.903	0.146	0.902	0.903	0.902	0.945	

=== Confusion Matrix ===

a	b	<-- classified as
6431	395	a = Normal
564	2471	b = Low

Experiment #9 using J48 Algorithm

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -C 0.5 -M 10
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves : 776
Size of the tree : 971

Time taken to build model: 0.46 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	8335	84.5249 %
Incorrectly Classified Instances	1526	15.4751 %
Kappa statistic	0.6253	
Mean absolute error	0.2052	
Root mean squared error	0.3379	
Relative absolute error	48.1468 %	
Root relative squared error	73.1959 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.913	0.307	0.87	0.913	0.891	0.895	Normal
	0.693	0.087	0.78	0.693	0.734	0.895	Low
Weighted Avg.	0.845	0.239	0.842	0.845	0.843	0.895	

=== Confusion Matrix ===

a	b	<-- classified as
6231	595	a = Normal
931	2104	b = Low

Experiment #10 using J48 Algorithm

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -U -M 2
Relation: Birth Weight-

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 unpruned tree

Number of Leaves : 2601
Size of the tree : 3312

Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	9155	92.8405 %
Incorrectly Classified Instances	706	7.1595 %
Kappa statistic	0.8255	
Mean absolute error	0.0823	
Root mean squared error	0.2478	
Relative absolute error	19.9715 %	
Root relative squared error	54.5918 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.953	0.131	0.947	0.953	0.95	0.949	Normal
	0.869	0.047	0.883	0.869	0.876	0.949	Low
Weighted Avg.	0.928	0.107	0.928	0.928	0.928	0.949	

=== Confusion Matrix ===

a	b	<-- classified as
6667	331	a = Normal
375	2488	b = Low

Experiment #11 using J48 Algorithm

=== Run information ===

Scheme: Weka.classifiers.trees.J48 -U -M 2
Relation: Birth Weight-
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 unpruned tree

Number of Leaves : 2601
Size of the tree : 3312

Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	9155	92.8405 %
Incorrectly Classified Instances	706	7.1595 %
Kappa statistic	0.8255	
Mean absolute error	0.0823	
Root mean squared error	0.2478	
Relative absolute error	19.9715 %	
Root relative squared error	54.5918 %	
Total Number of Instances	9861	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.953	0.131	0.947	0.953	0.95	0.949	Normal
	0.869	0.047	0.883	0.869	0.876	0.949	Low
Weighted Avg.	0.928	0.107	0.928	0.928	0.928	0.949	

=== Confusion Matrix ===

a	b	<-- classified as
6667	331	a = Normal
375	2488	b = Low

Annex B: the result of run information of second experiments using reduced attributes

Experiment #1 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Birth Weight-

Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   1309
Size of the tree :        1764

Time taken to build model: 0.75 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        9120               92.4855 %
Incorrectly Classified Instances      741                7.5145 %
Kappa statistic                        0.8198
Mean absolute error                    0.1023
Root mean squared error                0.2524
Relative absolute error               24.0022 %
Root relative squared error           54.6748 %
Total Number of Instances             9861

        === Detailed Accuracy By Class ===
                   TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                     0.963     0.16       0.931     0.963     0.947      0.954      Normal
                     0.84      0.037      0.909     0.84      0.873      0.954      Low
Weighted Avg.    0.925     0.123      0.924     0.925     0.924      0.954
=== Confusion Matrix ===
   a    b   <-- classified as
 6572  254 |    a = Normal
  487 2548 |    b = Low
```

Experiment #2 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.25 -M 5
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   929
Size of the tree :        1235
Time taken to build model: 0.47 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        8718               88.4089 %
Incorrectly Classified Instances      1143               11.5911 %
Kappa statistic                          0.7204
Mean absolute error                      0.1585
Root mean squared error                  0.3015
Relative absolute error                 37.202  %
Root relative squared error             65.3257 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.938     0.237      0.899     0.938     0.918      0.927    Normal
                 0.763     0.062      0.845     0.763     0.802      0.927    Low
Weighted Avg.    0.884     0.183      0.882     0.884     0.882      0.927

=== Confusion Matrix ===
   a    b   <-- classified as
 6402  424 |    a = Normal
  719 2316 |    b = Low
```

Experiment #3 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.25 -M 10
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   566
Size of the tree :        751

Time taken to build model: 0.5 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8190               83.0545 %
Incorrectly Classified Instances      1671               16.9455 %
Kappa statistic                          0.585
Mean absolute error                      0.2327
Root mean squared error                  0.3557
Relative absolute error                 54.606  %
Root relative squared error             77.0598 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.911     0.35       0.854     0.911     0.882      0.87     Normal
                 0.65      0.089      0.764     0.65      0.702      0.87     Low
Weighted Avg.    0.831     0.27       0.826     0.831     0.826      0.87

=== Confusion Matrix ===
   a    b   <-- classified as
 6218  608 |    a = Normal
 1063 1972 |    b = Low
```

Experiment #4 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.3 -M 2
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   1350
Size of the tree :         1820

Time taken to build model: 0.56 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        9163               92.9216 %
Incorrectly Classified Instances       698                7.0784 %
Kappa statistic                          0.8309
Mean absolute error                      0.0956
Root mean squared error                  0.2448
Relative absolute error                 22.4373 %
Root relative squared error             53.0294 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.963     0.147      0.937     0.963     0.95       0.958    Normal
                 0.853     0.037      0.911     0.853     0.881      0.958    Low
Weighted Avg.    0.929     0.113      0.929     0.929     0.929      0.958

=== Confusion Matrix ===
   a    b   <-- classified as
 6573  253 |    a = Normal
  445 2590 |    b = Low
```

Experiment #5 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.3 -M 5
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   968
Size of the tree :        1284

Time taken to build model: 0.52 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8749               88.7233 %
Incorrectly Classified Instances      1112               11.2767 %
Kappa statistic                          0.7286
Mean absolute error                      0.1545
Root mean squared error                  0.2975
Relative absolute error                 36.2506 %
Root relative squared error             64.4529 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.938     0.228      0.903     0.938     0.92       0.93     Normal
                 0.772     0.062      0.848     0.772     0.808      0.93     Low
Weighted Avg.    0.887     0.177      0.886     0.887     0.886      0.93

=== Confusion Matrix ===
   a    b   <-- classified as
 6406  420 |    a = Normal
  692 2343 |    b = Low
```

Experiment #6 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.3 -M 10
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   607
Size of the tree :    803

Time taken to build model: 0.44 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        8226               83.4195 %
Incorrectly Classified Instances      1635               16.5805 %
Kappa statistic                          0.5956
Mean absolute error                      0.2287
Root mean squared error                  0.3524
Relative absolute error                 53.6789 %
Root relative squared error             76.3439 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.91      0.337      0.859      0.91       0.884      0.875    Normal
                   0.663     0.09       0.767      0.663      0.711      0.875    Low
Weighted Avg.    0.834     0.261      0.83       0.834      0.831      0.875

=== Confusion Matrix ===
   a    b   <-- classified as
 6215  611 |    a = Normal
 1024 2011 |    b = Low
```

Experiment #7 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.5 -M 2
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   1384
Size of the tree :        1866

Time taken to build model: 0.55 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9197               93.2664 %
Incorrectly Classified Instances       664                6.7336 %
Kappa statistic                          0.8398
Mean absolute error                      0.0881
Root mean squared error                  0.238
Relative absolute error                 20.666  %
Root relative squared error             51.5713 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.962     0.133      0.942     0.962     0.952      0.962    Normal
                 0.867     0.038      0.91      0.867     0.888      0.962    Low
Weighted Avg.    0.933     0.104      0.932     0.933     0.932      0.962

=== Confusion Matrix ===
   a    b   <-- classified as
 6567  259 |    a = Normal
  405 2630 |    b = Low
```

Experiment #8 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.5 -M 5
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   1026
Size of the tree :        1351

Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8765               88.8855 %
Incorrectly Classified Instances      1096               11.1145 %
Kappa statistic                          0.7338
Mean absolute error                      0.1486
Root mean squared error                  0.2944
Relative absolute error                 34.8623 %
Root relative squared error             63.776  %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.936     0.217      0.906     0.936     0.921      0.934    Normal
                 0.783     0.064      0.845     0.783     0.813      0.934    Low
Weighted Avg.    0.889     0.17       0.887     0.889     0.888      0.934

=== Confusion Matrix ===

   a    b   <-- classified as
 6389  437 |    a = Normal
  659 2376 |    b = Low
```

Experiment #9 using J48 Algorithm with reduced attributes

```
=== Run information ===
Scheme:       Weka.classifiers.trees.J48 -C 0.5 -M 10
Relation:     Birth Weight-
Attributes:   11
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

Number of Leaves  :   668
Size of the tree :    876
Time taken to build model: 0.45 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        8241               83.5716 %
Incorrectly Classified Instances      1620               16.4284 %
Kappa statistic                          0.6018
Mean absolute error                      0.2229
Root mean squared error                  0.3491
Relative absolute error                 52.3095 %
Root relative squared error             75.6357 %
Total Number of Instances             9861

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.907    0.324      0.863     0.907     0.884      0.882    Normal
                   0.676    0.093      0.763     0.676     0.717      0.882    Low
Weighted Avg.    0.836     0.253      0.832     0.836     0.833      0.882

=== Confusion Matrix ===
   a    b   <-- classified as
 6189  637 |    a = Normal
  983 2052 |    b = Low
```

Annex C: the result of run information using PART algorithm all attributes

Experiment using PART Algorithm with all attributes

```
=== Run information ===
Scheme:       Weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     Birth Weight
Attributes:   15
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===

PART decision list
------------------

Number of Rules  :     460
Time taken to build model: 9.9 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        9304               94.3515 %
Incorrectly Classified Instances       557                5.6485 %
Kappa statistic                          0.8663
Mean absolute error                      0.0737
Root mean squared error                  0.2261
Relative absolute error                 17.3066 %
Root relative squared error             48.9915 %
Total Number of Instances             9861
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.966    0.107      0.953     0.966     0.959      0.961    Normal
                 0.893    0.034      0.921     0.893     0.907      0.961    Low
Weighted Avg.    0.944    0.084      0.943     0.944     0.943      0.961
=== Confusion Matrix ===
   a    b   <-- classified as
 6593  233 |    a = Normal
  324 2711 |    b = Low
```