ADDIS ABABAUNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND SCHOOL OF PUBLIC HEALTH

Predicting Nutritional Status of Women of Reproductive Age in Ethiopia

A Project submitted to the School of Graduate studies in partial fulfillment of the requirements for the degree of Master of Science in Health Informatics

By

Tigist Beyene

June, 2015

Addis Ababa, Ethiopia

ADDIS ABABAUNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND SCHOOL OF PUBLIC HEALTH

Predicting Nutritional Status of Women of Reproductive Age in Ethiopia

A Project submitted to the School of Graduate studies in partial fulfillment of the requirements for the degree of Master of Science in Health Informatics

By

Tigist Beyene

June, 2015

Addis Ababa, Ethiopia

Addis Ababa University

School of Graduate Studies

School of Information Science and School of Public Health

Predicting Nutritional Status of Women of Reproductive Age in Ethiopia

By

Tigist Beyene

Approved by the Examining Board

Name	Signature	Date
_____	_____	_____
Advisor		
_____	_____	_____
Advisor		
_____	_____	_____
Examiner		
_____	_____	_____
Examiner		

Declaration

I declare that this research project is my original work and it has not been presented for a degree in any other University. All the material sources used in this study are duly acknowledged.

_____

Tigist Beyene

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

_____

Dr. Million Meshesha

_____

Dr. Bilal Shikur

Acknowledgement

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

Secondly, I would like to forward a very much grateful thank to my advisors Dr. Million Meshesha and Dr. Bilal Shikur for their constructive comments and guidance while working on my thesis.

I would also like to thank Ato Birrara Ayele who is the FMOH Nutrition focal person, Ato. Birrara, helped me he selection of attributes related to my study topic and selected the best rule based on the algorithm generated.

I am very much grateful to Dr. Girmay Medhin to assist me morally and his unreserved advice and constructive comments from the beginning to the end of my study.

Table of Contents

Lists of Tables

Page

Lists of Figure

Page

ACRONMYS

ANN	Artificial Neural Network
EDHS	Ethiopia Demographic and Health Survey
DM	Data Mining
GUI	Graphical User Interface
KDD	knowledge discovery in databases
MIS	Management Information System
NNS	National Nutrition Strategy
PART	Partial decision tree
MLP	Multilayer Perception
ROC	Receiver Operator
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
WHO	World Health Organization

Abstract

Background: Nutritional status is the result of complex interaction between food consumption and the overall status of health and health care practices. Women of the reproductive age (15-49) are amongst the most vulnerable to risk of malnutrition. In 2011, among adults, 500 million women were anemic, and 500 million women were obese.

Objective: The main objective of this study is to apply data mining technique for constructing a predictive model that helps to predict nutritional status of women of reproductive age in Ethiopia.

Method: This study used a Hybrid data mining model and the dataset was extracted from the most Ethiopian Demographic and Health survey. To run the experiment used 18875 records, eighteen predicting variables and one outcome variable. Classification mining techniques are selected to build the model. Because of the their popularity in recent works J48, decision tree, PART rule induction, MLP Neutral Network, SMO support vector machine and naïve bayes algorithms were used as they implemented in Weka 3.6.10 tool.

Result: The best classification result, and a better predictive accuracy of women nutritional status was obtained from the unpruned PART rule induction. The experiment generated a model with accuracy of 83.14%, weighted precision of 83.1% and Weighted ROC area of 90.7%. Women's age, socioeconomic status, educational level sources of drinking water, latrine facility, breast feeding status, occupation, contraceptive method being under use, marital status, anemia level, residence and region, are the determinant factors of women's malnutrition.

Conclusion: The current result indicated that data mining is advantageous in bringing relevant information from large and complex dataset which can used for decision making. Program implements might also use the finding to identify women which needs special attention to reduce malnutrition.

CHAPTER ONE
INRODUCTION

1.1. Background

In a broader sense nutrition is defined as the science of foods, the nutrients and other substances their action, interaction, and balance in relationship to health and diseases, the process by which the organism ingests, digests, absorbs, transport and utilizes nutrients and dispose off their end products. Malnutrition is an impairment of health either from a deficiency or excess [1].

Ethiopia has the highest rates of malnutrition compared to other Sub-Saharan Africa country. The prevalence of malnutrition imposes significant costs on the Ethiopian economy as well as society [2]. The high mortality due to malnutrition leads to the loss of the economic potential of the women. Until recently, the broad multi-sectoral factors contributing to malnutrition had been insufficiently emphasized, with the focus placed on addressing food security as the primary means to address nutritional insecurity. To address these, the National Nutrition Strategy (NNS) was formulated and launched in 2008 [2].

The main goal of the Ethiopia NNS is to ensure that all Ethiopians attain adequate nutritional status in a sustainable manner, which is an essential requirement for a healthy and productive life [2]. During the implementation of the NNS, it was predicted that by the year 2015, certain nutrition milestones would be realized in Ethiopia towards achievement of the MDGs [2].

Nutrition of women affects a wide range of health and social issues, including family care and households food security [3]. The double burden of under nutrition and obesity is one of the leading causes of death and disability globally. In 2011, among adults, 500 million women were anemic, and 500 million women were obese [3].

The Ethiopian government is committed to accelerating implementation of the multisectorally harmonized National Nutrition Program to make a strong impact on nutrition and on overall wellbeing of the nation [4]. The National Nutrition program is designed to address both long-term and short-term nutrition goals in Ethiopia. The program outlines to plan for a package of proven, cost-effective nutrition interventions that will break the cycle of malnutrition [4].

Different demographic and socioeconomic factors may affect nutritional status of women of reproductive age in Ethiopia. To understand these factors there is a need to apply a data mining technology..Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both valid understandable and useful to the data owner [5]. It is also a process of discovering various models, summaries, and derived values from a given collection of data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful and interesting in that they lead to some advantage, usually an economic one [6]. It has a potential to identifying hidden knowledge from huge datasets [7]. Data Mining has the major tasks: predictive and descriptive modeling.

1.2. Statement of the problem

Malnutrition in all its forms is closely linked, either directly or indirectly, to major causes of death and disability worldwide.

Nutritional status is the result of complex interactions between food consumption and the overall status of health and health care practices. For women, improving overall nutritional status throughout the life cycle is crucial to maternal health. Women who become malnourished during pregnancy and children who fail to grow and develop normally due to malnutrition at any time during their life, including foetal development stage, are at increased risk of prenatal problems, increased susceptibility to infections, slowed recovery from illness, and possibly death [3].

The poor nutritional status of women has been a serious problem in Ethiopia for many years. Usually nutritional status of women is assessed by use of two anthropometric indices—height and body mass index (BMI) [8]. To derive those indices, the 2011 Ethiopia Demographic Health Survey EDHS measured the height and weight of women age 15-49 years [8]. Short stature reflects previous poor socioeconomic conditions and inadequate nutrition during childhood and adolescence. In a woman short stature is a risk factor for poor birth outcomes and obstetric complications [8]. For example, short stature is associated with small pelvic size, which increases the likelihood of difficulty during delivery and the risk of bearing low birth weight babies. A woman is considered to be at risk if her height is below 145 cm [8].

According to EDHS 2011, report 3 percent of Ethiopian women are below 145 cm in height. In general, height differs little with background characteristics. Women of short stature are most likely to reside in the Amhara region, and to have no education or primary education [8].

Adolescents (age 15-19) are more likely to be underweight (36 percent) than older women. Rural women also are more likely to be underweight than urban women, and those residing in the Afar region are the most likely to be underweight compared to other region. By education, women who have attended up to primary school have the greatest likelihood of being underweight. Those in the lowest wealth quintile also are more likely to be underweight than women in other wealth quintiles [8].

Overweight or obesity (BMI 25 kg/m2 or above) is not common among women in Ethiopia. Five percent are overweight (BMI 25-29 kg/m2), and just 1 percent are obese (BMI 30 kg/m2 or above). Younger women are less likely than older women to be overweight or obese. For example, 2 percent of women aged 15-19 are overweight or obese compared with 9 percent of women aged 40-49 years. Urban women are more likely to be overweight or obese (15 percent) compared to rural women (3 percent). In Addis Ababa and the Dire Dawa administrations overweight or obese is (20 percent and 19 percent, respectively), whereas, 3 percent of women in Benishangul-Gumuz and Tigray are overweight. Being overweight or obese is positively correlated with educational attainment; the proportion of overweight or obese women increases steadily from 4 percent among those with no education to 17 percent among those with more than secondary schooling. Similarly, the proportion of overweight or obese women increases as wealth increases, from 2 percent in the lowest wealth quintile to 16 percent in the highest quintile [8].

Women particularly in the child bearing age are especially vulnerable to malnutrition. Affluent women in this age group can also have poor nutritional status. However certain cultural and social practices may influence on the women nutrition and there is a scarce information about determinant factors of overweight/obese and underweight. Therefore, the purpose of this study is predicting women nutrition status that enable as to determine the determinant factors of nutrition. And there is no any research about nutritional status of women by using data mining technology in Ethiopia. Because of the above problems:

The study attempts to answer the following questions;

- What are the determinant factors that influence to the nutritional status of women?
- Which mining algorithm more suitable to build the nutritional status of women predictive model?

1.3. Objectives of the study

1.3.1. General Objective

The general Objective of this study is to construct a predictive model for determining the nutritional status of women of reproductive age in Ethiopia using data mining technology.

1.3.2. Specific objectives

To achieve the general objective of this study, the following specific objectives are formulated

1. To prepare the data for analysis and model building by cleaning, extracting and transforming in to a format suitable for the selected data mining algorithm.
2. To identify the most important attributes used to predict the nutritional status of women of reproductive age.
3. To identify factors that are associated with the nutritional status of women of reproductive age in each region of Ethiopia and to create a model that can be used to predict the nutritional status of women.
4. To evaluate the performance of the predictor model based on which the best model is selected for the prediction of the nutritional status of women.
5. To develop a graphically user friendly prototype system (interface) of the model to ease usage of the predictor model by domain users.

1.4. Scope and limitation of the Study

The scope of this study is to apply data mining techniques to predict nutritional status of women of reproductive age (i.e. 15-49) who have participated in EDHS 2011. The findings of this study as pertinent to the policy makers to make precise decision regarding the nutritional status of women. In this study classification algorithms of data mining are considered to create a prediction model so as to know the nutritional status of women.

The limitation of this study is lack of the relevant related literature on the data mining in this area

was one of the limitation in countered for experience sharing and comparing the result of this study.

1.5. Significance of the study

Since malnutrition in women is serious problem in Ethiopia, action to improve women's nutritional status should be foreseen at any time in the reproductive age of women. This study is therefore done to analyze the EDHS 2011 data in identifying some of the basic determinants of malnutrition among women in Ethiopia. Thus the findings of this study is used to predict nutritional status of women of reproductive age in Ethiopia. This study has to contributes a great deal of benefit to Policy makers and programmers so as make use of the model to develop new guidelines and policies and/or modify the existing ones in order to improve achievement of woman's nutrition programs goals in the country.

1.6. Methodology

1.6.1. Study Design

In this study Hybrid Data Mining (Hybrid-DM) process model is followed.[9]. This model combines best features of Cross-Industry Standard process for Data Mining (CRISP_DM) and Knowledge discovery in Database (KDD) methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives [9].

As shown in figure 1.1, Hybrid data mining model is a six step knowledge discovery process model such as understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge [9].

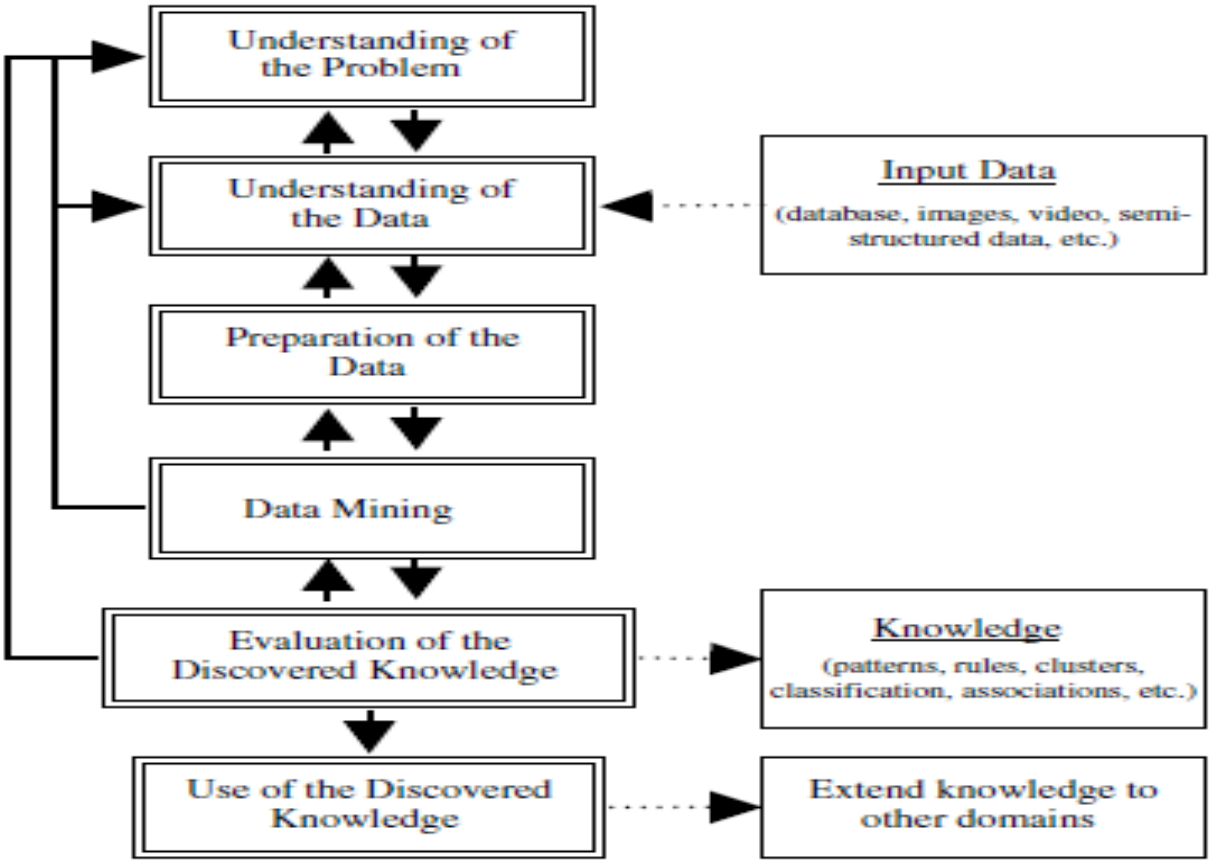


FIGURE 1.1: Hybrid-DM Process Models [9]

1.6.1.1. Understanding of the problem domain

The model was needed to predict which region would most likely affected by women malnutrition to give emphasis before it causes further problem to the society. For understanding the problem domain of women nutritional status in Ethiopia, This study was done in consultation with the domain experts and also read related literature on the problem to get familiar with the business, to know the specific terminology to determine the study objectives, and to define problems.

1.6.1.2. Understanding of the data

In this step data was collected from the original source from EDHS 2011. The data was in Statistical Package for Social Science (SPSS) it contains a total of 11654 records. Most of the dataset contains Nominal values. Descriptive summarization and visualizations of the data conducted using SPPS. Data was checked for completeness, redundancy, missing values,

plausibility of attribute values. The step includes verification of the usefulness of the data with respect to the DM goals. Thus, for this study, total amount of 10875 dataset utilized. This data was partitioned and used for training and testing the accuracy of the model.

1.6.1.3. Preparation of the data

This step is one of the crucial step to produce data used for modeling by Waikato Environment for Knowledge analysis (Weka) software. It is concerned with deciding which data should be used as an input for DM methods in the subsequent step. It involves data clearing, attribute selection, and transformation. In order to correct errors identified trough observation form the preprocessing stage, measures like filling in missing values based on the idea of observing neighboring records, correcting inconsistencies like spelling error checking the completeness of records, and removing instances with missing values were undertaken. SMOTE was applied to overcome the problem of high imbalance in the class value. Finally the dataset ready for the data mining process contains only 19 attributes and 18875 records.

1.6.1.4. Data mining

To build a predictive model from the cleaned data, WEKA data mining software was used. WEKA is a tool containing numerous machine learning algorithms that can be applied to achieve the objective of this study. It supports several standard data mining tasks. More specifically, data preprocessing, clustering, classification, regression, visualization and feature selection and also this software is platform independent. Decision tree, rule induction, support vector machine, Artificial Neural Network, and Naïve bayes algorithm were used among acceptance of recent study [10]. With this reason this study has employed five classification DM algorithms to develop and compare the classification model.

Accordingly, among the available algorithms in WEKA machine learning software; J48, PART, MLP, SMO, and Naïve bayes algorithms were applied on the EDHS 2011 data to come up with the predictive model for predicting nutritional status of women.

1.6.1.5. Evaluation of the discovered knowledge

The result of knowledge discovery process was evaluated to reach at a certain conclusion which is relevant to the problem at hand. Evaluating the discovered knowledge also comprises

understanding the results, and checking the discovered knowledge is novel and interesting. In this study confusion matrix and accuracy, sensitivity, specificity and precision to evaluate the performance of each of the model. The patterns were checked for their interestingness together with the domain area experts interpretations were critically commented by experts.

1.6.1.6. Uses of the discovered knowledge

This is a final step of knowledge discovery process which consists of planning where and how to use the discovered knowledge. This last step determines the success of the entire knowledge discovery process. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented.

- A prototype is developed to the discovered knowledge using Visual Basic 2010 programming langue that enables end users to predict nutritional status of women.
- A softcopy of this study will be uploaded to Addis Ababa university electronic resource official website
- A hard copy of the documentation will be available at the bibliographic library of the school of information science and library of the school of public health.
- Maximum effort will be exerted to publish the result on different journals to initiate other interested groups to find gaps and do more research in the area.

1.7. Ethical Consideration

The current study has taken ethical consideration three major ethical issues in the course of the study. The first one is the study does not use personal identification like name during mode building or in reporting its findings. The second on e is the findings of this work are not to be used for commercial gain Moreover; the findings of the study would not a .harm study subject. The ethical clearance was obtained from school of public health and a letter of cooperation was obtained from school of information science before the launch of data collection.

CHAPTER TWO
LITRATURE REVIEW

Data mining emerged during the late 1980s, made great steps during the 1990s, and continues to grow into the new millennium [9]. Data mining has attracted a great deal of attention in the information industry and the society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [9]. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [9].

2.1. Overview of Data Mining

Data Mining is a multidisciplinary field, drawing work from areas including database management systems, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing, and data visualization [11]. The rules include the iterative process of detecting and extracting patterns from large databases [11].

There are a number of definitions of Data Mining in the literature [11]. However, they all have in common the following; extraction, knowledge, and large data. It refers to the extract on or “mining” hidden knowledge from large amount of data. The process of performing data analysis may uncover important data patterns. The exploration and analysis, by automatic or semiautomatic means large quantities of data in order to discover meaningful patterns and rules [11].

Data mining is a process of extracting and identifying useful information and subsequent hidden knowledge from large databases and data warehouses using statistical, mathematical, artificial intelligence and machine learning technique. Data mining applies modern statistical and computational technologies in its quest to expose useful pattern hidden within the large databases [5].

2.2. Data mining tasks

The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data [10]. The tasks of data mining can be modeled as either Predictive or Descriptive in nature. A Predictive model makes a prediction about values of data using known results found from different data, while the Descriptive model identifies patterns or relationships in data [7].Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. Predictive model data mining tasks include classification, prediction, regression and time series analysis. The Descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis [7].

Among Predictive models, Classification is probably the most widely used data mining approaches [7]. According to Siraj and Abdoula [7] the three common characteristics of classification tasks are:

• Learning is supervised

• The dependent variable is categorical

• The model built is able to assign new data to one of a set of well-defined classes [5].

Unlike a classification model, the purpose of Prediction model is to determine the future outcome rather than current behavior.

Another Predictive model known as statistical Regression is a supervised learning technique that involves analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the development of a model that can predict these attribute values for new cases. Prediction applications with one or more time-dependent attributes are called time-series problems. Time series analysis usually involves predicting numeric outcomes such as the future price of individual stock [10].

The second approach of data mining is known as Descriptive method. Descriptive data mining is normally used to generate frequency, cross tabulation and correlation. Descriptive method can be defined to discover interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data. Summarization maps data into subsets with associated simple Descriptions. Basic statistics such as Mean, Standard Deviation, Variance, Mode and Median can

be used as Summarization approach. In Clustering, a set of data items is partitioned into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar. Associations or Link Analysis are used to discover relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. Association Rules is a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. The investigation of relationships between items over a period of time is also often referred to as Sequence Analysis. Sequence Analysis is used to determine sequential patterns in data. The patterns in the dataset are based on time sequence of actions, and they are similar to association data, however the relationship is based on time [7].

2.3. Classification algorithms

A classification technique is a systematic approach to building classification models from an input data set. There are several models developed for classifying high dimensional data, such as decision tree, Rule Induction, neural networks, support vector machine and Naive Bayes and so on. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute sets and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability that accurately predicts the class labels of previously unknown records [11].

2.4. Data Mining Methodologies

In the latest years, the growth and consolidation of the data mining has been coming into occurrence. Some efforts are being done that seek the establishment of standards in the area, both by academics and by people in the industry field. The academics efforts are centered in the attempt to formulate a general framework for DM [12]. The efforts in the industrial field concern mainly the definition of processes or methodologies that can guide the implementation of DM applications. To implement DM applications, one needs to select the best modeling technique. So in this study, relative comparison of, KDD, SEMMA, CRISP-DM and Hybrid-DM are made to select the most suitable modeling [12].

2.4.1. Knowledge Discovery in Database (KDD)

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database [11].

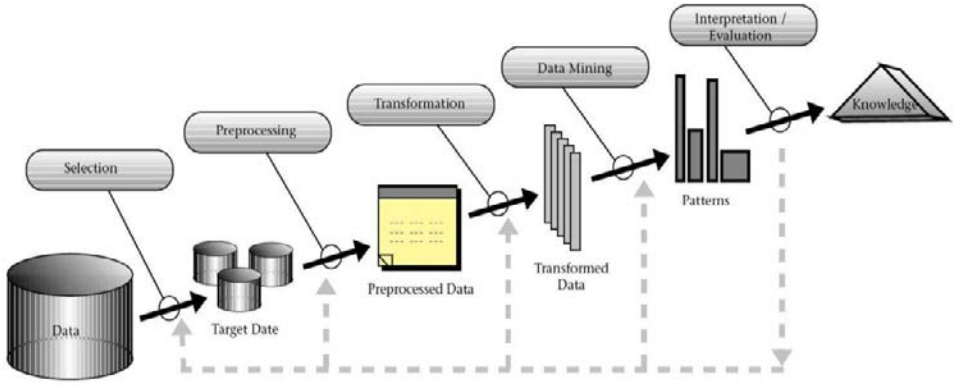


Figure 2.1: KDD- Process Model [11]

KDD is a five stage process, these are: selection, preprocessing, Transformation, Data mining and Interpretation/Evaluation

The ***Selection*** stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed**.** The ***Preprocessing*** consists of the target data cleaning and pre processing in order to obtain consistent data. Further, ***Transformation*** consists of the transformation of the data using dimensionality reduction or transformation methods .Once the dataset is prepared the next step is ***Data Mining*** which is consists of the searching process for identification of patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction). Finally, the constructed model is evaluated. The ***Interpretation/Evaluation*** is consists of the interpretation and evaluation of the mined patterns [10, 11].

The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. Additionally, the KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It must also be continued by the knowledge consolidation and incorporating this knowledge into the system [10, 11].

2.4.2. Sample, Explore, Modify, Model and Assess (SEMMA)

The SEMMA process was developed by the Statistical Analysis System (SAS) Institute [20]. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a DM project. The SAS Institute considers a cycle with five stages for the process. The SEMMA analysis cycle guides the analyst through the process of exploring the data using visual and statistical techniques, transforming data to uncover the most significant predictive variables, modeling the variables to predict outcomes, and assessing the model by testing it with new data [12].

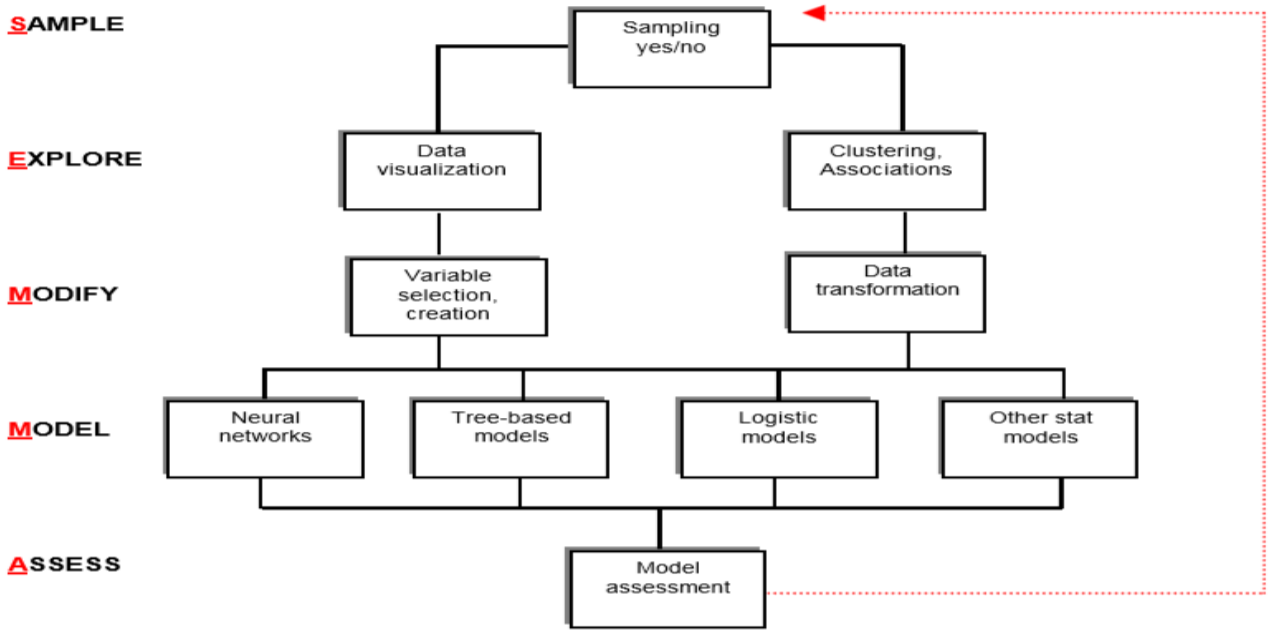


Figure 2.2: Step in SEMMA process: [12]

As shown in figure 2.2 ***Sample*** is the first step which create one or more data tables by sampling data from the data warehouse. Mining a representative sample instead of the entire volume radically reduces the processing time required to obtain business information. After sampling the data the next step is ***Explore*** the data visually or numerically for trends or groupings. Exploration helps

to refine the discovery process and techniques such as factor analysis, correlation analysis and clustering. The third step is ***modifying*** the data which refers to creating, selecting, and transforming one or more variables to focus the model selection process in a particular direction, or to modify the data for clarity or consistence. The forth step is creating a data ***model*** which involves using the DM software to search automatically for combination of data that predicts the desired outcome reliably. The last step is to ***assess*** the model to determine how well it performs. A common means of assessing a model is to set aside a portion of the data during the sampling stage. If the model is valid it should work for both the reserved sample and for the sample that was used to develop the model [12]

Although the SEMMA process is independent from DM chosen tool, it is linked to the SAS enterprise miner software and pretends to guide the user on the implementations of DM applications. SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for this conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals [12].

2.4.3. Cross-Industry Standard Process for Data Mining

Cross-Industry Standard Process for DM (CRISP-DM) is a knowledge discovery approach which is widely used by industry members [12]. This model consists of six phases intended as a cyclical process [12].

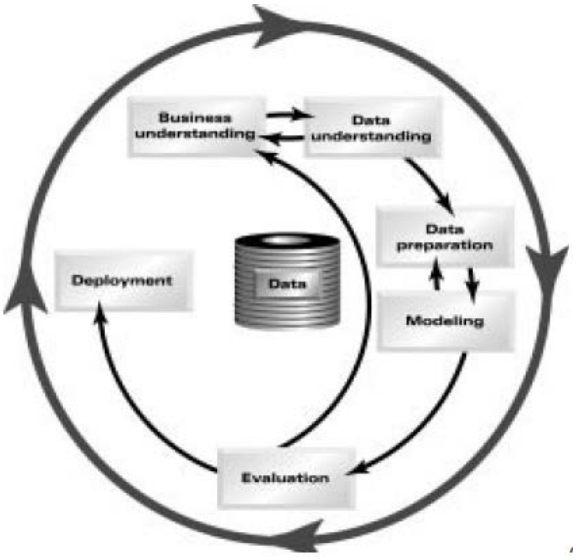


FIGURE : 2.3: The CRISP-DM Model

Business understanding includes determining business objectives, assessing the current situation, establishing data mining goals, and developing the project plan. Once the business objectives and project plan are established, the ***data understanding*** starts with an initial data collection and proceeds with activities in order to get familiarity with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to/from hypotheses for hidden information. After the resource data available are identified, the data ***preparation*** covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools. After preparation of data various ***modeling*** techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary. At the ***Evaluation*** stage in the project you have built a model (or models) that appear to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the DM results should be reached. The final stage is ***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. It often involves applying "live" models within an organization's decision making processes [12],

2.4.4 Hybrid Data mining Process Model

Hybrid-DM model is developed by Cios et al [9] on the CRISP-DM model by adopting it to academic research. It is a six stage process modeling which constitutes; understanding of the problem domain, understanding of the data, preparation of data, data mining and evaluation of the discovered knowledge and use of the discovered knowledge.

2.5. Comparison of data mining methodology

The comparison among the four methodologies is done based on the steps contained, applicability to academic researchers in general and its applicability to the identified problem. The first comparison is done by comparing KDD and SEMMA DM methodologies in the steps contained to govern the data mining task. By doing a comparison of the KDD and SEMMA stages we would affirm that they are equivalent:

- Sample can be identified with selection,
- Explore can be identified with the Pre processing
- Modify can be identified with Transformation
- Model can be identified with DM
- Assess can be identified with Interpretation/Evaluation

Examining it thoroughly, we may affirm that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process, since it is directly linked to the SAS enterprise miner software. There is no phase which can let this study to get familiar with the business in order to have initial insights about the problem and a stage to communicate the findings with the domain users.

Comparing the KDD AND CRISP-DM stages is not as straight forward as in the SEMMA situation. Nevertheless, we can first of all observe that the CRISP-DM methodology incorporates the steps that as referred above, must precede and follow the KDD process that is to say:

- The business understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user.
- The deployment phase can be indentified with the consolidation by incorporating this knowledge in to the system.

Concerning the remaining stages, we can say that:

- The data understanding phase can be identified as the combination of selection and preprocessing
- The data preparation phase can be identified with Transformation
- The modeling phase can be identified with data mining
- The evaluation phase can be identified with interpretation/evaluation

This comparison shows that CRISP-DM is better than KDD as well as SEMMA including additional stages for business understanding and deploying the final outcome or discovered knowledge in to the existing system.

For the comparison of CRISP-DM and Hybrid-DM they are almost equivalent with the development stages contained but with some adjustments done for Hybrid-DM. The main differences and extensions with CRISP-DM are Hybrid-DM:

- Provides more general, research-oriented description of the steps,
- Introduced a data mining step instead of the modeling step,
- Introduced several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the Hybrid model has more detailed feedback mechanisms) and
- Modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in the other domains.

Considering the above details on both CRISP-DM and Hybrid-DM process modeling, the one which suits this study is Hybrid-DM modeling due to its nature to be made for academic environments and the opportunity of using the discovered knowledge in to other domains. The other important reason for the selection of Hybrid-DM modeling is, it is initially tested with many health related works.

Table 2.1 Summary of data mining methodologies

KDD	SAMMA	CRISP	Hybrid
Pre KDD	……..	Business Understanding	Understanding of the problem
Selection	Sample	Data Understanding	Understanding of the data
Preprocessing	Explore		
Transformation	Modify	Data Preparation	Preparation of the Data
Data Mining	Model	Modeling	Data Mining
Interpretation Evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
Post KDD	……..	Deployment	Use of the Discovered knowledge

2.6. Data mining tools

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow the choice of the most suitable tools becomes increasingly difficult [14]. Nowadays, it is very often possible to look for many kinds of data mining tools both freely and commercially. In this study the main concern are the five well known open source mining software tools. Namely, Rapid Miner, Weka, Rattle, KNIME, Orange.

Weka, Rapid Miner and KNIME are developed in the Java software language 1[5]. Rattle is fully R-based application and Orange is integrated with python. Among all open source data mining tools, Weka and Rapid Miner have the biggest and most active user communities. Both of them quickly implement and integrate new and emerging machine learning algorithms into their systems [15].

Weka, as one of the best-known open-source data mining software tools, has an impressive array of data mining components, which have, in fact, been integrated into many other data mining tools including Rapid Miner, Rattle, and KNIME. Weka consists of four major applications: Explorer for exploring data, Experimenter for performing experiments and conducting statistical tests between learning schemes, Knowledge Flow for incremental learning, and Simple CLI (Command Line Interface) to allow direct execution of Weka commands. For beginners, it is best for them to start with weka Explorer, as it provides a relatively simple and easy to learn interface to access weka data mining components [15].

Rapid Miner (formerly YALE) is build on top of Weka and includes additional powerful data analysis functions such as data preprocessing, Visualization, and additional machine learning algorithms. In addition, its user interface is more intuitive than Weka knowledge Flow. Users with limited knowledge in computer science and programming may find Rapid Miners learning curve to be substantial [15].

KNIME has one of the best built-in online support features, which is very helpful for new users who are in the process of building their data mining workflows. KNIME also support running R and python scripts. Another nice feature of KNIME is its integration of the Chemistry

Development Kit with additional nodes for the processing of chemical structures, compounds, etc [15].

Rattle provides a graphical user interface (GUI) specifically for R data mining user. Although an understanding of R is not required to begin using Rattle for basic data mining functions, Rattle is particularly suited for users familiar with R. In addition, Rattle integrates two sophisticated tools for interactive graphical data analysis [15].

Orange has a very simple and intuitive graphical interface (GUI) for users with limited knowledge in data mining. Compared to the other data mining tools, its strength is its interactive visualization function, Which enables users to set visualization parameters and choose data points or nodes directly from a graph [15].

2.7. Overview of Nutrition

Nutrition is the science of food values. It is relatively a new science, which was evolved from chemistry and physiology. The effect of food in our body is explained in nutrition. In other words, nutrition are defined as food at work in the body. In addition nutrition must be concerned with the social, economic, cultural and psychological implication of food and eating [2]. Nutrients are defined as the constituents of food, which perform important functions in our body. If these nutrients are not present in our food in sufficient amount, the result is ill health. Important nutrients include carbohydrates, proteins lipids, vitamins, minerals, and water. Malnutrition is an impairment of health either from a deficiency or excess or imbalance of nutrients. In other words, malnutrition refers to both under nutrition and over nutrition [1].

2.8. Data Mining In Health Care

Healthcare generates mountains of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce huge of clinical data. This data is a strategic resource for healthcare institutions [16].

The use of data mining applications in healthcare is the realization that data mining can generate information that is very useful to all parties involved in the healthcare industry. Data mining applications can also benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices [16].

There is vast potential for data mining applications in healthcare. Generally, these can be grouped as the evaluation of treatment effectiveness; management of healthcare; customer relationship management; and detection of fraud and abuse [17].

To aid healthcare management, data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims [17].

2.9. Related Work

One study was conducted in Ethiopia [18], on Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: The source data employed for this research purpose is 2011 EDHS dataset. This dataset is collected from 2006/2007-2010/2011. The attributes selected for the prediction purpose, were from the five years survey data. These attributes are, Mother's age, Mother's educational level, Mother's Body Mass Index (BMI), Mother's occupation, Residence, Region, Wealth quintile, Size of child at birth, child's age, child's sex, child's Height for age Z- score (HAZ), child's weight for age Z- score (WAZ), child's Weight for Height Z-score (WHZ), anemia level, total number of children and ever had vaccinated.

The study followed by hybrid methodology of Knowledge Discovery Process (KDP). To achieve the goal building predictive model using data mining techniques. Classification algorithms using a dataset of 9,607 records.

In this study six experiments have been conducted using three data mining classification algorithms i.e. J48 algorithm, Naïve Bayes and PART rule induction classifier in order to build a model that predicts nutritional status of under-five years children in Ethiopia. In this study, the model created using PART pruned rule induction classifier registers good performance and hence selected for further analysis/rule tracing.

Another work conducted in India, [19] attempt the identification of malnutrition with use of supervised data mining techniques-decision tree and artificial neural networks. The sources of data employed for this research National Family Health Survey (NFHS) III conducted by Indian

institute of population science. The attribute selected for this research are Anemia level, Breast feed Cereals, Grain food, Fruits, Green vegetables, Fish food/ meat, egg and tinned powder/milk.

The study used classification data mining technique which is decision tree and artificial neural network. The researcher used Iterative Dichotomies 3 (ID3), random forest tree and Multilayer perception (MLP) algorithms to provide the nutritional status of children age under five. In this study the model created using a multilayer perception (MLP) neural network classifier and random forest tree register good performance.

2.10. Factors Associated With women's nutritional Status in Ethiopia

The study which has used data from 2000, 2005 and 2011 nationally representative of Ethiopian Demographic and Health surveys [20] showed that the prevalence of overweight/obesity among women in Addis Ababa increased significantly by 28%; while underweight decreased by 21% between 2000 and 2011. Specifically, the prevalence of urban obesity increased by 43.3% i.e., from 3.0% to 43.3% in about 15 years. Overall, more than one-third (34.7%) of women in Addis Ababa were either under or overweight. Women's age and proxies for high socio-economic status (i.e. household wealth quintile, educational attainment, access to improved source of drinking water, and television watching) were positively associated with being overweight.

In another hospital based study [21] targeting nutritional status of adults living with HIV/AIDS at the University of Gondar referral hospital northwest Ethiopia. It conducted from Oct- 30 2007. The prevalence of malnutrition (BMI<18.5kg/m2) was 27.8%. The percentage of body weight lost was (BWL>5%) was 60.9% and severe malnutrition (BWL>20%) accounted for 10.1%. Income, duration of ART in months, presence of eating problems, and nutritional support were significantly associated with malnutrition (BMI<18.5). BWL was significantly associated with nutritional support and duration of ART.

A study done by [22] identified determinants of nutritional status of Women and children in Ethiopia, the study based on data from the 2000 demographic and health survey with reference to 13447 women age 15-49 years and 9768 children under five interviewed mothers complete and plausible anthropometric data. The indicator used to assess chronic energy deficiency malnutrition in women is body mass index also known as the Quetelet index. the percentage of

stunted children was a bit higher among stunted mothers and normal height mothers 64.3% of the children of stunted mother were stunted, the level of stunting underweight and wasting was also higher in children malnourished mothers (BMI<18.5) as compared to well-nourished mothers (BMI >=18.5), no statistical significant association was observed. the authors conclude that the socioeconomic and demographic variables have significant influences on the odds of chronic energy deficiencies (CED) in women and malnutrition in children. Region of residence, household economic status, woman's employment status and decision making power over her income woman's age and marital status are important determinants of CED among reproductive age women.

Another study made by [23], the nutritional status of adolescent girls from rural community of Tigray, northern Ethiopia. The study was used for the data analysis anthropometric and socio demographic information from 211 adolescent girls representing 650 randomly selected households from thirteen communities in Tigray. The study compared the height for age and BMI for age with WHO growth reference. The cross-sectional prevalence of stunting and thinner were 26.5% and 58.3% respectively Lack of latrine facilities was significantly associated with stunting and thinness. Age was strong predictor of stunting and thinness.

CHAPTER THREE
Data Mining Algorithms

There are various DM techniques used their appropriateness to be applied in different health care areas. This study has incorporated classification algorithms to build the prediction model. Then J48 from decision tree, PART from rule induction, Naive bayes from bysian, sequential minimal optimization from support vector machine, and Multilayer perception form artificial neural network are selected to run the experiments due to their acceptance in recent research [24].

3.1 Decision Tree

Decision trees are supervised algorithms which recursively partition the data based on its attributes until some stopping condition is reached. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch depending from that node corresponds to one of the attributes values (outcome) of the test, and each leaf node holds a class label. The topmost node in a tree is the root node [25].

Decision Tree Classifier is one of the possible approaches to multistage decision-making. The most important feature of decision tree classifiers is their capability to break down a complex decision making process into a collection of simpler decisions, hence providing a solution, which is often easier to interpret. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.

During tree construction, attribute selection measures are used to select the attribute that best partitions the tuples into distinct classes. When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand [25].

J48 Decision tree

J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool for generating a pruned or unpruned C4.5 tree [25]. Pruning usually results in reducing size of tree, avoids unnecessary complexity, and to avoid over fitting of the data sets when classifying new data. Over fitting can lead an excessively large number of rules [26]. C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan [9]. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [25]. J48 builds decision trees from a set labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting he data into smaller subsets [9].

J48 examines the normalized information gain that results from choosing an attribute for splitting the data. To the decision, the attribute with the highest normalized information gain is used. Then the algorithms recur on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain [9]. The following algorithm shows the generation of a decision tree from a training tuples of data partition.

Partition D.
Input:
Data partition, D, which is a set of training tuples and their associated class labels;
attribute list, the set of candidate attributes;
attribute selection method, a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Output: A decision tree.
Method:
Create a node N;
If tuples in D are all of the same class, C then
Return N as a leaf node labeled with the class C

If attribute list is empty then
Return N as a leaf node labeled with the majority class in D; // majority voting
Apply attribute selection method (D, attribute list) to find the “best” splitting criterion:
Label node N with splitting criterion;
If splitting attribute is discrete-valued and
multi way splits allowed then // not restricted to binary tree
attribute list ← attribute list ←splitting attribute; // remove splitting attribute
for each outcome j of splitting criterion
// partition the tuples and grow sub trees for each partition
Let Dj be the set of data tuples in D satisfying outcome j; //a partition
If Dj is empty then
Attach a leaf labeled with the majority class in D to node N;
Else attach the node returned by generate decision tree (Dj, attribute list) to node N
End for
Return N; [9, 25]

J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation [9].

3.2. Rule Induction

Rule induction is a method for deriving a set of rules to classify cases. Although decision trees can produce a set of rules, rule induction methods generate a set of independent rules which do not necessarily form a tree [26]. Rule induction is one of the most important techniques of machine learning. Regularities hidden in data are frequently expressed in terms of rules; rule induction is one of the fundamental tools of data mining the same time [9].

Usually rules are expressions of the form,
If (attribute – 1; value - 1) and (attribute – 2; value -2)
And (attribute –n; value –n) then (decision; value) [9].

PART Rule Induction

PART (partial decision tree) is a rule induction algorithm which grabs rule from a decision tree. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees [6]. To generate such a tree, the conjunction and pruning operations are integrated in order to find stable sub tree that can be simplified no further [6]. Once this sub tree has been found, tree building ceases and a single rule s read off.

Initialize **E** to the instance set

For each class C, from smallest to largest

BUILD:

Split E into Growing and Pruning sets in the ratio 2:1

Repeat until (a) there are no more uncovered examples of C; or (b) the description length (DL) of rule set and examples is 64 bits greater than the

Smallest DL found so far, or (c) the error rate exceeds 50%:

GROW phase: Grow a rule by greedily adding conditions until the rule is 100% accurate by testing every possible value of each attribute and selecting

the condition with greatest information gain G

PRUNE phase: Prune conditions in last-to-first order. Continue as long as the worth W of the rule increases

OPTIMIZE:

GENERATE VARIANTS:

For each rule R for class C,

Split E afresh into Growing and Pruning sets

Remove all instances from the Pruning set that are covered by other rules for C

Use **GROW** and **PRUNE** to generate and prune two competing rules from the newly split data:

R1 is a new rule, rebuilt from scratch;

R2 is generated by greedily adding antecedents to R.

Prune using the metric A (instead of W) on this reduced data

SELECT REPRESENTATIVE:

Replace R by whichever of R, R1 and R2 has the smallest DL.

MOP UP:

If there are residual uncovered instances of class C, return to the BUILD stage to generate more rules based on these

CLEAN UP:

calculate DL for the whole rule set and for the rule set with each rule in turn omitted; delete any rule that increases the DL

Remove instances covered by the rules just generated

Continue [6]

3.3. Artificial Neural Networks Algorithm

An artificial neuron is a computational model inspired in natural neurons. Natural neurons receive signals through synapses located on the dendrites or membrane of the signals received are strong enough, the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons. The complexity of real neurons is highly abstracted when modeling artificial neurons [12, 27]. These basically consist of inputs, which are multiplied by weights, and then computed by a mathematical function which determines the activation of the neuron. Another function computes the output of the artificial neuron sometimes in dependence of certain threshold. ANN combines artificial neuron in order to process information [27].

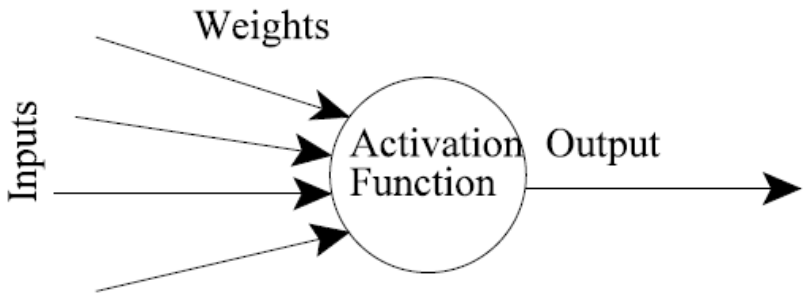


Figure 3.1: Artificial Neuron [27]

Neural Network technology uses a multilayered approach that approximates complex mathematical function to process data [10, 14]. Neural networks are trained that a particular input leads to a specific target output. Based on a comparison of the output and the target, the

network is trained until the network output matches the target. Typically many such input/target pairs are used to train a network [10].

Multilayer feed forward Neural Network

The backpropation algorithm performs learning on a multilayer feed-forward neural network. It learns a set of weights for prediction of the class label of tuples. Multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer [12, 25].

Each layer is made up of units. The inputs to the network corresponding to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of neuron like units, known as a hidden layer [20]. The output of the hidden layer units can be input to another hidden layer, and so on. Then number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples [12].

The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer. It is fully connected in that each unit provides input to each unit in the next forward layer. Multilayer feed-forward neural networks are able to model the class prediction as a nonlinear combination of the inputs [25].

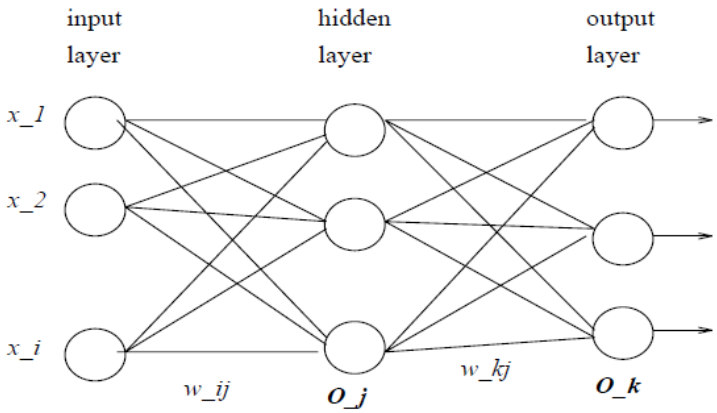


Figure 3.2: A multilayer feed-forward neural network [12]

To compute the net input to the unit, each input connected to the unit [12]

$$(Ij) = \sum_i W_{ij} O_i + \theta_j$$

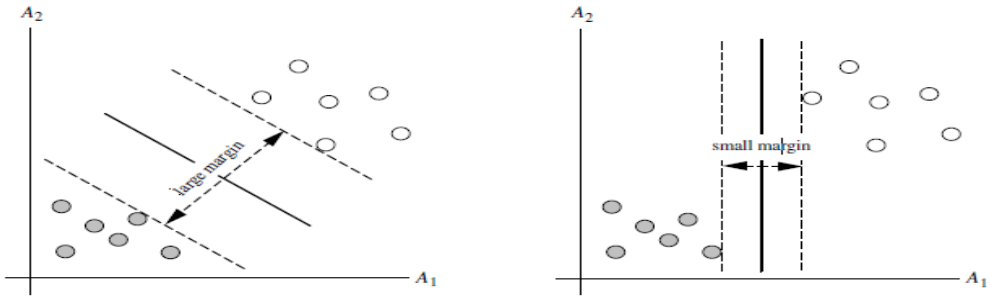
where Wj is the weight of the connection from unit I in the previous layer to unit j; Oi is the output of unit I from the previous layer; and θ_j is the bias of units which as acts as a threshold in that serves to vary the activity of the unit.

Each unit in the hidden and output layers takes its net input, and then applies an activation function. The output of the activation function symbolizes the activation of the neuron represented by the unit. The logistic or sigmoid function is used. Given the net input I_j to unit j, then O_j the output of unit j, is computed as:

$$Oj = 1/1+e^{-1}$$

3.4. Support Vector Machine (SVM) algorithm

Support vector machine is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high [28]. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data [28]. SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two dataset. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyper-plane, which are pushed up against the two data sets [29]. A good separation is achieved by the hyper-plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. This hyper-plane is found by using the support vectors and margins [29].



(a) (b)

Figure 3.3: possible separating hyper planes [12]

From these associated margins, the one with the larger margin should have greater generalization accuracy. Thus, any point that lies above the separating hyper plane satisfies

$$w_o+w_1x_1+ w_2x_2 > 0$$

Similarly, any point that lies below the separating hyper plane satisfies

$$w_o+w_1x_1+ w_2x_2 > 0$$

Sequential Minimal Optimization (SMO)

The Sequential minimal –optimization algorithm implements for training a support vector classifier, using kernel functions such as polynomial or Gaussian kernels. Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default, the coefficients in the output are based on the normalized data. Normalization can be turned off or the input standardized to zero mean and unit variance [6].

3.5 Naive Bayes Algorithms

The naïve bayes classifier is a simple classifier based bayes' law with strong independence assumptions among features is comparable to other state-of-the-art classifiers, namely, ID3 decision tree, J48 decision tree, and c4.5 decision tree [30]. The algorithm assumes that the object's attributes are independent [31]. The naïve byes classifier often works very well in practice, and excellent classification results may be obtained even when the probability estimates contain large errors [31].

The Naïve Bayesian classifier makes the assumption of class conditional independence i.e., that given the class label computation. When the assumption holds true, then the naïve Bayesian classifier is most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables. Bayesian belief networks specify joint conditional probability distributions. They allow class conditional independencies to be performed. These networks are also known as belief networks, Bayesian networks, and probabilistic networks.

3.6 Performance evaluation for predictive modeling

In any branch of science, it is almost a common requirement that performance of various models have to be compared with each other to understand the suitability of a model to a given problem. In data mining also it is a common requirement. Once a predictive model is developed using the

historical data such as EDHS 2011 data, the model should have to be checked as to how well it performs for the approaching data that it has not seen during the model building process. In this study different classifiers has used for building the predictive model, and confusion matrix was used for evaluating the performance of the model.

Confusion matrix is a simple performance analysis tool typically used in supervised learning. It is used to represent the test result of a prediction model. Each column of the matrix represents the instances in predicted class, while each row represents the instances in the actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing the class with other that is commonly called misclassification.[9].

Confusion matrix shows, for which the various values and related equations are described. Few of these equations are very relevant for performance analysis [9].

Confusion matrix		**predicted**	
		Negative	**Positive**
Actual	**Negative**	**a**	**b**
	positive	**b**	**d**

Figure 3.4: Confusion matrix

The entries in the confusion matrix have the following meaning in the context of a data mining problem:

- a- is the number of correct predictions that an instance is negative,
- b- is the number of incorrect predictions that an instance is positive,
- c- is the number of incorrect of predictions that an instance negative,
- d- is the number of correct predictions that an instance is positive.

Several standard terms are defined for the two class matrix such as accuracy, recall, precision [25].

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation [25]:

$$AC=(a+b)/(a+b+c+d)$$

The recall or true positive (TP) rate is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP=d/(c+d)$$

The false positive (FP) rate is the proportion of negative cases that were incorrectly classified as positive, as calculated using the equation:

$$FP=b(/a+b)$$

The true negative (TN) rate is defined as the proportion of negative cases that were classified correctly, as calculated using the equation:

$$TN=a/(a+b)$$

The false negative (FN) rate is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN=c/(c+b)$$

Precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P=d/(b+d)$$

The concepts of these two class problem can be extended to a multi class problem by focusing one of the classes as positive at a time and the rest as negative. The average of these parameters like precision, recall etc. for individual classes becomes the final values of the entire model.

Receiver operator characteristic test (ROC) is a plot of the true positive rate against the false positive rate for the different possible cut points of a diagnostic test.

A ROC curve demonstrates several things[25]:

1. It shows the tradeoffs between sensitivity and specificity , any increase in sensitivity will be accompanied by a decrease in specificity.
2. The closer the curve follows the left hand border and then the top border of the ROC space, the more accurate the test.

3. The closer the curve comes to the 45 degree diagonal of the ROC space. The less accurate the test.
4. The area under the curve is a measure of text accuracy.

ROC graphs are two-dimensional graphs in which true positive (TP) rate which is sensitivity is plotted on the Y axis and false positive (FP) rate which is specificity is plotted on the X axis [30].

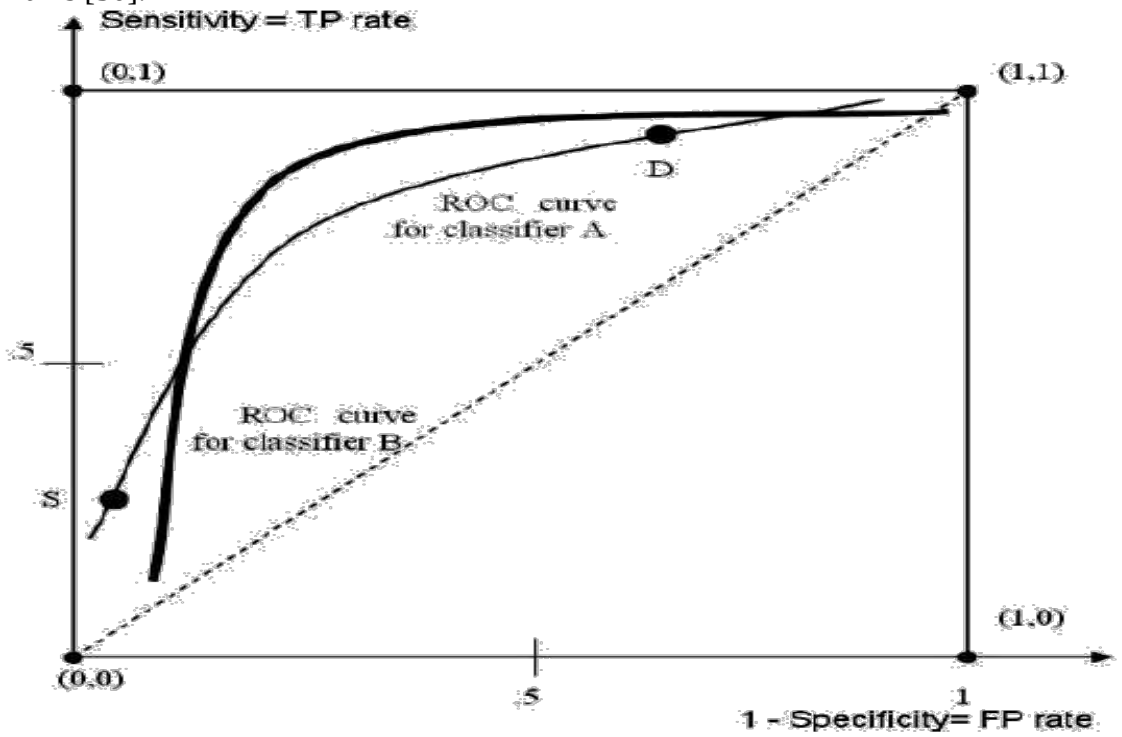


Figure 3.5: ROC curve for two classifiers [25]

In order to decide which of th**e** classifier in figure 3.6 constitutes a better model/classifier of the data, visual analysis could be performed, that is the curve more to the upper left would indicate a better classifier. However, the curves often overlap, as shown on the above figure for two classifiers, in this case the popular method called Area under curve (AUC) is used. Since the AUC is a proportion of the area of the unit square, its value will always be between 0 and 1.0. This method chooses a classifier that has maximum area under its corresponding ROC curve; the larger the area, the better performing the model/classifier is.

CHAPTER FOUR

Preparation of the data

DM is a technology that uses various techniques to discover hidden knowledge from data stored in large databases, data warehouses and other massive information repositories [12]. To discover non- trivial knowledge and patterns, the database must undergo effective data preparation to bring a valid output. In this study a hybrid data mining model is used, which is a six steps knowledge discovery process. Among these business understanding, data understanding and data preprocessing are meant to prepare the data for data mining tasks.

4.1. Business understanding

Business understanding is the key element of a data mining study to know what the study is for. This begins with a managerial need for new knowledge, and an expression of the business objective regarding the study to be undertaken. The main outputs of Business understanding phase are the definition of business and data mining objectives as well as business and data mining evaluation criteria. Business understanding is mainly concerned with the identifying business objectives and determination of data mining goals.

4.1.1. Identifying business objectives

Nutritional status is the result of complex interactions between food consumption and the overall status of health and health care practices. For women, improving overall nutritional status throughout the life cycle is crucial to maternal health. Some evidence In developing countries indicate individuals that is, women with body mass index (BMI) below 18.5 show a progressive increase in mortality rates as well as increased risk of illness. For social and biological reasons, women of the reproductive age are amongst the most vulnerable to malnutrition. Increased prenatal and neonatal mortality, a higher risk of low birth weight babies, still births, and miscarriage are some of the consequences of malnutrition in women [22]. Women who are overweight are more likely to have chronic conditions such as diabetes or hypertension which could impact the health of their pregnancies. Also overweight/obese is also a risk factor for infertility. Women with a body mass index (BMI) of 28 or above are approximately 2.5 times more likely to have ovulatory infertility than women with a BMI of 18 to 22 [27]. Women with

a higher re-pregnancy BMI are at increased risk for gestational diabetes, preeclampsia/eclampsia, abnormal labor, premature delivery and cesarean delivery [27],

4.1.2. Determination of data mining goals

The DM goals are derived from the business objectives so that the mining tasks were done targeting the stated goals. Accordingly the following DM goals were identified to guide upcoming events like experimentation and prototype development. The first goal is given the demographic and socio economic data, predict women nutritional status and the second goal is from the identified predicting variables, determine those having a better prediction performance.

4.2. Data Understanding

Since data mining is task-oriented, different business tasks require different data set. Therefore appropriate data is an important factor to achieved data mining project goals. And data understanding phase focuses on creating a target dataset with selected set of the attribute that is relevant to the process. Without understanding the existing it is difficult to draw the target data set from the original since the world data is unclean and not appropriate at the source to run mining process [6]. The original dataset is entered in SPSS, from this statistical tool the data exported to EXCEL file to convert to .CSV format which is accepted by Weka data mining tool.

4.2.1 Data Source and data collection

The data source for this study has taken from Ethiopia Demographic and Health Survey (EDHS) 2011 dataset and nutritional status of women of reproductive age data**.** The data was collected information from all women age 15-49. The collected data is the anthropometric measures to assess the nutritional status of women and Conduct hemoglobin testing on women of the same age to provide information on the prevalence of anemia.

4.2.2 Description of Data

The EDHS survey is conducted in five years interval since 2000. The 2011 EDHS is conducted by Central Statistics Agency under the support of Ministry of health. The primary objectives of the 2011 EDHS are to provide up-to-date information for planning, policy formulation, monitoring and evaluation of population and health programs in the country. The survey was conducted at the beginning of the last term of the Millennium Development Goal (MDG) plan

period and to provide data for the assessment of MDGs. The 2011 survey provides critical information for the use as baseline data in monitoring and evaluation of the growth and Transformation Plan (GTP) as well as various sector development policies and programs.

In the 2011 EDHS collected information on population and health condition which covers family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women's empowerment and knowledge of HIV/AIDS.

The survey is nationally representative survey of 11654 records (instances) on Women of reproductive age in order to classify nutrition status, men age of 15-59, and under five year children on 920 variables (attributes) .This sample provides estimates of health and demographic indicators at the national and regional levels, and for rural and urban areas.

From original women's data other attributes and unrelated attributes with nutritional status of reproductive age of women removed. Finally the selection of the dataset is performed by the help of literature and domain experts.

The whole attributes in the original data set is not concerned for this Experimentation. Thus only the relevant attributes are considered so as to achieve the objective of the study. From total of 290 attributes which were found in women of reproductive age records, 19 variables were selected, which are 18 predictive or Independent and 1 dependent variables. The detail is given in tale below:

Table 4.1: Selected attribute with their description from EDHS Data Set

No	Variable Name	Description	Data Type	Value
1	Age	Women Age Group	Categorical	1=15-19, 2=20-24, 3=25-29, 4=30-34, 5=35-39, 6=40-44, 7=45-49
2	Marital Status	Current marital status of women	Nominal	1=Never in union, 2=Married, 3=Living with partner 4=divorced, 5=no longer living together/separated
3	Educational	Women level	Nominal	0=No education, 1=primary, 2=secondary, 3=

	level	of education		Technical Vocational, 4=Higher
4	Wealth Index	Wealth Index	Nominal	1=Poorest, 2=Poorer, 3=Middle, 4=richer, 5=richest
5	Residence	Palce of residence	Nominal	1-urban 2=rural
6	Region	11administrative region of Ethiopia	Nominal	1=Tigray, 2=Affar, 3=Amhara, 4=Oromiya, 5=Somali, 6=Benishangul-Gumuz, 7=SNNP, 12=Gambela, 13=Harari, 14=Addis Ababa, 15 Dire Dawa
7	Drinking water	Sources of drinking water	Nominal	11=piped into dwelling, 12=piped to yard/plot, 13=public tap/standpipe, 21=tube well or borehole, 31=protected well, 32=unprotected well, 41=protected spring, 42=unprotected spring 43=river/dam/lake/ponds/stream/canal/irrigation channel, 51=rain water, 61= tanker truck, 62=cart with small tank, 71=bottled water, 96=other,
8	Frequency of listening radio	Frequency of listening radio	Nominal	0=not at all, 1=less than once a week, 2=at least once a week
9	occupation	Women currently working	Nominal	0=No, 1=Yes
10	Relationship hh head	Women relationship HH head	Nominal	1=head, 2=wife, 3=daughter, 4=daughter in law, 5=granddaughter, 6=mother, 7=mother in law, 8=sister, 9=co-spouse, 10=other relative, 11= adopted/foster child/step child, 12=not related, 13=Niece, 98=Don't know
11	Contraceptive	Current contraceptive method	Nominal	0=not using, 1=pill, 2=IUD, 3=injections, 4=diaphragm, 5=condom, 6female sterilization, 7=male sterilization, 8=periodic abstinence, 9=withdrawal, 10=other,

				11=implants/norplant,12=abstinence, 13=lactational amenorrhea, 14=female condom, 15=foam or jelly, 17= other modern method, 18=standard days method,
12	Anemia	Women Anemia level	Nominal	1=severe, 2=moderate, 3=mild, 4=not anemic
13	Pregnancy	Currently pregnant	Nominal	0=No or unsure, 1=Yes
14	Breastfeeding	Currently breastfeeding	Nominal	0=no 1=yes
15	Religion	Religion	Nominal	1=Orthodox, 2=catholic, 3=protestant, 4=Muslim, 5=traditional, 6=other
16	Parity	Number of living children	Categorical	1=No child, 2=1-2, 3=3-4, 4=5-6, 5=7-8, 6=9-10, 7=11-12
17	Toilet facility	Type of Toilet facility	Nominal	11=Flush to piped sewer system, 12=flush to septic tank, 13=Flush to latrine, 14=flush to somewhere else, 15=flush don't know where, 21=ventilated improve pit latrine (VIP), 22=pit latrine with slab, 23=pit latrine without slab/open pit, 31=no facility/bush/field, 41=composting toilet, 42=bucket toilet, 43=hanging toilet/latrine, 96=other
18	Sex of HH	Sex of Household head	Nominal	1=Male 2=Female
Dependent (outcome) variable				
19	BMI	Body Mass Index	Nominal	Underweight, Normal and overweight/obese

Body Mass Index (BMI) is an anthropometric index of weight and height (stature) that is defined as body weight in kilogram divided by height in meter squared [33]. Like weight-for-height, BMI is a screening tool used to identify individuals who are underweight or overweight [33].

BMI is an indicator of total body fat and is therefore an indicator of health risk BMI is used by healthcare professionals to screen for overweight and obese individual. This information can then be used by healthcare professionals to assess a patient's health risk. A high BMI is a risk factor for disease and even death [34].

Women nutritional status can be categorized in to three, Underweight, Normal and overweight,

Underweight: Women Body mass Index (BMI) is less than 18.5. on the other hand

Normal: women body mass index (BMI) is 18.5-24.9.

Over weight/obese: women body mass index (BMI) is greater than 25.

4.3. Data Preprocessing

Data preprocessing is an important step which ensures the data quality and improves the efficiency and ease of mining process.

Real world data sets are usually not directly suitable for performing Data Mining algorithms. They contain incomplete, noisy, and may be inconsistent in addition real world data set tend to be too large and high dimensional. This is due to the fact that incomplete data lacking attribute values or certain attributes of interest, or containing only aggregate data, noisy containing errors, or outlier values which deviate from the expected, and inconsistent containing discrepancies in codes or names.

Data preprocessing includes data cleaning to remove noise and outliers, data integration to integrate data obtained from multiple information sources, data reduction to reduce the dimensionality and complexity of the data and transformation to convert the data into suitable forms for mining. Core mining refers to the essential process where various algorithms are applied to perform the data mining tasks. The purpose of data preprocessing is to clean selected data for better quality.

Data preparation generates a data set smaller than the original one, which can significantly improve the efficiency of Data Mining. It generates quality data, which leads to quality patterns.

For instance we can recover incomplete data by filling the values missed, or by reducing ambiguity. And we can purify data errors, or remove outliers the other is resolving data conflicts using domain knowledge or expert decision to settle discrepancy. Good data preparation is key to producing valid and reliable models [25].

While data mining technology can support the data-analysis applications to identify nutrition status of women of reproductive age rate, it must be possible to prepare quality data from the raw data so as to enable quality knowledge discovery from the given data [25].

Different data preprocessing tasks were involved in this study such as relevant attributes selection, data reduction, instance selection, and data cleaning. The other task is carried out by data transformation such as normalization and aggregation. Data discretization also another task of data preprocessing, which replaces numerical values with nominal. Data reduction is meant to reduce the volume but producing similar analytical result.

4.3.1. Exploratory Data Analysis

Effective data mining needs exploratory data analysis. Descriptive data summarization techniques can be used to classify the representative properties of data and highlight which data values should be treated as noise or outliers. Moreover missing values can easily be detected using exploratory data analysis [12, 25].

To understand the nature of the data values in the selected EDHS 2011 dataset, an exploratory data analysis was done so as to explore by categorical and nominal attributes, to expose the valid and missing instances and the percentage.

Table 4.2: Frequency Distribution of Socio Demographic Characteristics

Characteristics	Number	Percent
Residence		
Urban	1986	17.0
Rural	9668	83.0
Total	11654	100%
Marital status		
Never in union	71	0.6
Married	10190	87.4
Living with Partner	589	5.1
Widowed	213	1.8
Divorced	402	3.4
No longer living together/separated	169	1.6
Total	11654	100.0
Educational level		
No Education	8142	69.9
Primary	2030	25.1
Secondary	386	3.3
Higher	101	0.9
Technical vocational	95	0.8
Total	11654	100.0
Woman age group		
15-19	516	4.4
20-24	2342	20.1
25-29	3642	31.3
30-34	2369	20.3
35-39	1793	15.4
40-44	738	6.3
45-49	254	2.2
Total	11654	100.0
Currently Working		
Yes	3487	29.9
No	8150	69.9
Total	11637	99.9
Missing 9	17	0.1
Total	11654	100.0
Religion		
Orthodox	3617	31.0
Catholic	109	0.9
Protestant	2237	19.2
Muslim	5447	46.7
Traditional	96	0.8
Other	143	1.2
Total	11649	100
Missing 99	5	0
Total	11654	100.0

Table 4.2 shows most attributes have no missing values except religion and women current working have missing values**.**

Table 4.3: Frequency distribution of Region

Variable	Number	Present
Region		
Tigray	1202	10.3
Affar	1130	9.7
Amhara	1294	11.1
Oromiya	1761	15.1
Somali	1027	8.8
Benishangul-Gumuz	1020	8.8
SNNP	1614	13.8
Gambela	851	7.3
Harari	659	5.7
Addis Ababa	400	3.4
Dire Dawa	696	6.0
Total	11654	100.0

As shown in table 4.3 attribute is a nominal variable with 11 values there was no missing values found in the dataset**.** Oromiya region is most frequent, where as the least frequent is Addis Ababa region.

Table 4.4: Frequency distribution of HH head gender and Women's Relationship to HH

Variables	Number	Percent
Household head gender		
Male	9469	81.3
Female	2185	18.7
Total	11654	100.0
Relationship to HH)		
Head	1704	14.6
Wife	8766	72.2
Daughter	662	5.7
Daughter- in- law	162	1.4
Grand daughter	24	0.2
Other relative	88	0.8
Sister	91	0.8
Niece	46	0.4
Adopted/faster child/step child	16	0.1
Mother	5	0.0
Mother –in-law	6	0.1
Not related	84	0.7
Total	11654	100.0

Table 4.4 shows the attribute is also nominal variable with no missing values in the data. And the most frequent values is found in the women relationship to HH is wife. Male HH head are more frequent.

Table 4.5: Frequency Distribution of Anemia level, pregnancy, breastfeeding, and contraceptive method

Variable	Number	Percent
Anemia level		
Not anemic	8366	71.8
Moderate	731	6.3
Mild	1942	16.7
Severe	140	1.2
Total	11179	95.9
Missing 9	475	4.1
Total	11654	100.0
Currently pregnant		
Yes	1203	11.2
No unsure	10351	88.8
Total	11654	100.0
Currently breastfeeding		
Yes	7785	66.8
No	3869	33.2
Total	11654	100.0
Uses of contraceptive method		
Not using	9229	79.2
Pill	219	1.9
IUD	39	0.3
Injection	1706	14.6
Condom	25	0.2
Female sterilization	16	0.1
Periodic abstinence	79	0.7
Withdrawal	19	0.2
Other	11	0.1
Implants/Norplant	309	2.7
Lactation amenorrhea (LAM)	2	0.0
Total	11654	100

As presented table 4.5 the attribute anemia level is a nominal value with most of the observation are not anemic, It has missing values. The un pregnant women is more frequent and there is no missing values. the attribute currently breast feeding is a nominal value without missing values and the lactating women are more frequent. The un using women is more frequent in attribute of use of contraceptive method.

Table 4.6: Frequency distribution of Wealth status and listening radio

Variables		Number	Percent
Wealth index			
	Poorest	3625	31.1
	Poorer	2114	18.1
	Middle	1872	16.1
	Richer	1870	16.0
	Richest	2173	18.6
	Total	11654	100
Frequency of listening radio			
	At least once a week	1799	15.4
	Less than once a week	3233	27.7
	Not at all	6613	56.7
	Total	11654	100.0

Table 4.6 shows the attributes are nominal and there are no missing values in the two attributes data. The poorest is more frequent in the wealth index attribute, not listening radio at all are more frequent.

Table 4.7: Frequency distribution of sources of drinking water

Variable	Number	Percent
Sources of drinking water		
Piped into dwelling	66	0.6
Piped yard/plot	726	6.2
Public tap/standpipe	2180	18.7
Tube well or borehole	659	5.7
Protected well	1209	10.4
Unprotected well	677	5.8
Protected spring	801	6.9
River/dam/lake/ponds/stream/canal/ irrigation channel	2396	20.6
Rain water	105	0.9
Tanker track	83	0.7
Cart with small tank	137	1.2
Bottled water	2	0
Other	47	0.4
Not a dejure resident	269	2.3
Total	11649	100
Missing 99	5	0

Table 4.7 shows the attribute of drinking water is nominal, and there is a missing value. River, dam, lake, ponds, steam, canal irrigation channel source are most frequent.

Table 4.8: Frequency distribution of toilet facility

Variable	Number	Percent
Toilet facility		
Flush to piped sewer system	60	0.5
Flush to septic tank	72	0.6
Flush to pit latrine	157	1.3
Flush to somewhere else	4	0.0
Flush don't know where	8	0.1
Ventilated improved pit latrine (VIP)	192	1.6
Pit latrine with slab	988	8.5
Pit latrine without slab/open pit	3612	31.0
No facility/bush/field	5877	50.4
Compositing toilet	386	3.3
Bucket toilet	2	0.0
Hanging toilet/latrine	8	0.1
Other	17	0.1
Not a dejure resident	269	2.3
Total	11652	100
Missing 99	2	0.0
Total	11654	100.0

Table 4.7 shows the attribute of toilet facility is nominal, and there is a missing value. No facility is most frequent.

4.3.2. Data Cleaning

Data cleaning is a time consuming and labor intensive procedure, but it is absolutely necessary for successful data mining. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. In general data cleaning decreases errors and increases the data quality.

4.3.3. Handling missing values

In large database, there is a problem of missing values. A missing value may have been accidentally not computerized, or purposely not obtained for technical, economic or ethical reasons [31].

Some of the potential methods that can be used to handle missing value are described below [17]

- Deleting the missing attribute or deleting a record with missing values
- Substitute missing values with most likely values
- Replace one missing value with all possible values for that attribute
- Replacing the missing value with mean of the attribute for numerical attributes or the mode for categorical variables
- Replacing the missing values with a value generated at random from the variable distribution observed

Fortunately, as presented in table 4.9 below attributes have missing values. These attributes are, religion, frequency of listening radio, anemia level, current working, Body Mass Index, toilet facility and sources of drinking water. The percentages of missing values in each of the mentioned attributes are less than 5%. Hence, it is assumed that percent of the missing values is not considerably large to the extent that it can significantly influence the final conclusion [36]. With this assumption, the analysis file was reduced by deleting all cases with the missing values on the above mentioned attributes [35].

Table 4.9: attributes missing value and percentage

Variable	Valid	Percent	Missing values	Percent
Religion	11637	99.9	17	0.1
Frequency of listening radio	11645	99.9	9	0.1
Anemia level	11179	95.9	475	4.1
Currently working	11637	99.9	17	0.1
Body mass index	11409	97.9	245	2.1
Toilet facility	11652	100	2	0.0
Sources of drinking water	11649	100	5	0.0

4.3.4. Data Transformation

Data transformation is about transforming the data to make it appropriate for mining. Therefore this study based on the dataset encoding data variable. Table 4.10 shows the data transformation result for the five attributes.

Table 4.10: Data Encoding

No	variable	Old value	New value	New Value
1	Wealth index	1=Poorest 2=Poorer 3=Middle 4=Richest5=Richer	Poor, Middle, Rich	1=Poor2=Middle3=Rich **Note** Poorest, Poorer change in to **poor** also Richest and Richer were changed in to **Rich**
2	Sources of Drinking water	11=piped into dwelling, 12=piped to yard/plot, 13=public tap/standpipe, 21=tube well or borehole, 31=protected well, 32=unprotected well, 41=protected spring, 42=unprotected spring, 43=river/dam/lake/ponds/stream/ canal/irrigation channel, 51=rain water, 61= tanker truck, 62=cart with small tank, 71=bottled water, 96=other,	Improved , Non Improved, other	piped into dwelling, piped to yard/plot, public tap/standpipe, tube well or borehole, protected well, protected spring. Rain water, and bottled water change in to **Improved,** unprotected well, unprotected spring, river/ dam/ lake/ ponds/ stream/ canal/irrigation channel, tanker truck, cart with small tank change in to **Non improved** , 96=other
3	Parity	0,1,2,3,4,5,6,7,8,9,10,11,12	0,1,2,3	0= no child 1=1-2, 2=3-4, 3=5+
4	Current contraceptive method	0=not using, 1=pill, 2=IUD, 3=injections, 4=diaphragm, 5=condom, 6female sterilization, 7=male sterilization, 8=periodic abstinence, 9=withdrawal, 10=other, 11=implants/Norplant, 12=abstinence, 13=lactational	Modern, traditional, other	0=not using, pill, IUD, injection, diaphragm, condom, female sterilization, male sterilization, implants/Norplant, lactation amenorrhea, female condom, foam or jelly, other modern method are changed **modern contraceptive** Periodic abstinence, withdrawal, other are changed traditional contraceptive

		amenorrhea, 14=female condom, 15=foam or jelly, 17= other modern method, 18=standard days method,		
5	Type of Toilet facility	11=Flush to piped sewer system, 12=flush to septic tank, 13=Flush to latrine, 14=flush to somewhere else, 15=flush don't know where, 21=ventilated improve pit latrine (VIP), 22=pit latrine with slab, 23=pit latrine without slab/open pit, 31=no facility/bush/field, 41=composting toilet, 42=bucket toilet, 43=hanging toilet/latrine, 96=other	Improved, non improved, other	Flush to piped sewer system, flush to septic tank, flush to pit latrine, ventilated improved pit latrine (VIP), pit latrine with slab, compositing toilet were changed **Improved .** pit latrine without slab/open pit, no facility/ bush/field, bucket toilet, hanging toilet/latrine were changed **Non improved** **96= other**

CHAPTER FIVE

Experimentation and Discussion

In this study five data mining algorithms were used to achieve the objective of the study. These mining algorithms are J48 Decision tree, Naive Bayes, PART Rule induction Sequential Minimal Optimization, support vector machine (simple Minimal Optimization), and Multilayer perception Artificial Neural Network. In the application of each of the five mining algorithms 10-fold cross validation technique was used to train and test classifiers.

5.1. Experimental Setup

The initial dataset had 920 attributes or variables and 11654 records. After preprocessing the data and deleted records with missing values greater than 50% on a given variable, the data was reduced to nineteen attributes (i.e. 18 independent and 1 outcome) and 10875 cases. The processed dataset was converted to Comma separated values (CSV) file format so as to make it ready for WEKA machine learning software. Moreover, these experiments are done by using WEKA version 3.6.10. The data is imbalanced if the classification categories are not approximately equally represented [6]. Performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large. In the case of women nutrition status data the class variable BMI has a higher imbalance. Therefore, the study used SMOTE automatic operations by filter where minority classes are oversampled by 75 generated synthetic examples of minority class and adding them to the dataset. This way, the class distribution in the dataset changes and probability of correctly classifying minority class increases.

-a- -b-

Figure 5.1: Weka view of side by side class variables on women's nutritional status. (Figure "a" is original data; figure "b" is balanced data using SMOTE).

As shown in figure 5.1 the SMOTE operation applied to the minority class. Originally there is 7336 cases in the majority class and only 654 cases in the minority class. After applying the SMOTE analysis the minority class comes to 5018.

The test option used by this study was 10 fold cross validation for partition of the dataset into training and test set. Data was divided into 10 folds and each in turn was used for testing and the remainder is used for training. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus, the learning procedure is executed 10 times on different training sets [6]. The study used 10 folds because it is commonly used data partitioning technique for training and testing a classifier. Extensive tests on numerous datasets, with different learning techniques, have shown that 10 is the right number of folds to get the best estimate of error and there is also some theoretical evidence that backs this up [6].

Selection of variables

One important feature of Weka which may be crucial for some learning schemes, is the opportunity of choosing a smaller subset from large set of attributes. Some algorithms work

slower when the attributes have lots of instances. Another reason could be that some variables might not be relevant. Determining the relevance of the variables is searching within the domain of all possible subsets of variables and finding the subset that works best for classifying. In the classification process two operators are needed, namely, subset evaluator and search method. The search method traverses that attribute subset space and used the evaluator for quality measure. Both of them can be chosen and configured same to the filters and classifiers [6].

CFS subset selection evaluator selects best attribute by evaluating the value of a subset of variable by considering the individual predictive attribute. CFS subset evaluator selected six best attributes, namely Region, Residence, toilet facility, Wealth status, currently pregnancy status, and age.

Table 5.1: Best attributes by CFS subset evaluator

Rank	Attribute Name	Data type
1	Region	Nominal
2	Residence	Nominal
3	Toilet facility	Nominal
4	Wealth status	Nominal
5	Currently pregnant	Nominal
6	Age	Nominal

5.2. Experimentations to build a predictive Model for Nutritional status Measured using (BMI)

Different experiments were conducted using five algorithms by changing the parameters contained in each of them. In these experiments Body Mass Index was the outcome variable, the independent variables were eighteen.

5.2.1. J48 decision tree

Four experiments are conducted using J48 decision tree by the parameter with default value and changing value of unpruned value in to True value with all and best selected attributes.

- Setting 1: J48 Experiment Pruned with all attributes
- Setting 2: J48 Experiment unpruned with all attributes
- Setting 3: J48 Experiments pruned with best selected attributes
- Setting 4: J48 Experiments unpruned with best selected attributes

Summary of the performance of the four experiments using J48 decision tree is summarized in table 5.2. In the first setting, 18 attributes and 18875 records were used by taking the default parameter value with pruned value. This experiment has generated a model with a tree size of 4999 and leaves sizes of 3836.

In the second setting, 18 attributes and 18875 records were used but there was a change of pruned into unpruned True to run the experiment. This experiment result resulted in a relatively larger tree having a size of 7669 and leaves of 5951 were generated.

In the third setting, six best selected attributes and 18875 records were used by taking the default parameter value with pruned. The experiment generated a model with a tree size of 651 and 470 leaves.

In the fourth setting six best selected attributes and 18875 records were used by taking unpruned value True to run the experiment. Relatively larger tree having a size of 802 and leaves of 575 were generated.

Table 5.2: J48 Experimental result using J48 decision tree

Experiments	**Accuracy**	**WTP rate**	**WFP rate**	**WTP precision**	**WTP recall**	**WF measure**	**W ROC area**
J48 Pruned with all attributes	80.286	0.803	0.107	0.802	0.803	0.802	0.899
Unpruned J48 True with all attributes	**81.917**	**0.819**	**0.098**	**0.818**	**0.819**	**0.818**	**0.908**
J48 Pruned with Best attributes	64.471	0.645	0.191	0.643	0.645	0.644	0.803
J48 un Pruned True with Best attributes	64.54	0.645	0.189	0.643	0.645	0.644	0.809

Result from J48 decision tree with all 18 attributes and 18875 records showed that the experiment has generated a model with accuracy of 80.28% weighted precision of 80.2% and weighted ROC area of 89.9%. In the second experiment the only change of pruned to un pruned true value, resulted in model with accuracy, weighted precision and weighted ROC area of 81.9%, 81.8% and 90.8% respectively.

In the third setting J48 pruned with the best selected attributes of result showed that the experiment has constructed a model with accuracy of 64.47%, Weighted precision of 64.3% and Weighted ROC area of 80.3%.The fourth setting used the selected best attributes and records, but due to the change of un pruned value to true, the experiment constructed a model with accuracy, weighted precision and weighted ROC area of 64.5%, 64.3% and 80.9% respectively.

Based on the above experiment, the un pruned J48 decision tree has scored a better accuracy than pruned J48. Therefore, a model constructed by the unpruned J48 decision tree using all attributes was selected after comparing with other classifiers generated under the above experiments.

5.2.2. PART rule induction

Four experiments are conducted using PART rule induction by the parameter with default value and changing value of un pruned value in to True value with all and best selected attributes.

- Setting 1: Experiment Pruned PART with all attributes
- Setting 2: Experiment unpruned PART with all attributes
- Setting 3: Experiments pruned PART with best selected attributes
- Setting 4: Experiments unpruned PART with best selected attributes

In the first setting which has used the 18 attributes and 18875 records by taking the default parameter value with pruned value has generated a model with 1364 rules.

In the second setting, the same number of attributes and records were used but there was a change of pruned into unpruned True to run the experiment. Relatively larger number of rules having 2500 rules was generated.

In the third setting of the experiment, six best selected attributes and 18875 records are used by taking the default parameter value with pruned. This experiment has generated a model with 265 rules.

In the fourth setting, the same records and attributes as that of 3rd setting were used by taking un pruned value True to run the experiment. Relatively larger number of rules having 578 rules was generated. Experimental results of these four experiments are presented in table 5.3.

Table 5.3: PART rule induction experiment result

Experiments	**Accuracy**	**WTP rate**	**WFP rate**	**WTP precision**	**WTP recall**	**WF measure**	**W ROC area**
PART with Pruned and all attributes	80.651	0.807	0.106	0.806	0.807	0.806	0.898
PART unPruned True with all attributes	**83.147**	**0.831**	0.092	**0.831**	**0.831**	**0.831**	**0.907**
PART Pruned with Best attributes	64.37	0.644	0.191	0.641	0.644	0.642	0.807
PART unPruned True with Best attributes	64.64	64.6	0.188	0.643	0.646	0.644	0,813

As shown in the above table, 18 attributes and 10875 records are used by taking the default value with PART pruned parameter value The result showed that the experiment has generated a model with accuracy of 80.65%, weighted precision of 80.6% and weighted ROC area 89.8% for the first setting. For the second setting the same number of attributes and records are used but due to the change of pruned to un pruned true value, experiment has constructed a model with accuracy, weighted precision and weighted ROC area 83.147%, 83.1.7% and 90.7% respectively.

In the third setting PART pruned with the best selected attributes of result showed that the experiment has constructed a model with accuracy of 64.37%, weighted precision of 64.1% and Weighted ROC area of 80.7%.The fourth setting used the same number of attributes and records, but due to the change of un pruned value to true, the experiment constructed a model with accuracy, weighted precision and weighted ROC area 64.64%, 64.3% and 81.3% respectively.

Based on experiment result, the unpruned PART with all attributes has scored a better accuracy than pruned PART. Therefore, the un pruned PART rule induction has been selected after comparing with other classifiers generated under this experiments.

5.2.3. Sequential Minimal Optimization (SMO) Experiments

There are two experiments conducted using SMO classification algorithm. The first experiment was carried out by taking the default parameter values with 18875 records and 18 attributes and second experiment was conducted by taking the default parameter value with the best the six selected attributes and 18875 records. Table 5.4 presents experimental result using SMO.

Table 5.4: Performance of SMO

Experiments	Accuracy	WTP rate	WFP rate	WTP precision	WTP recall	WF measure	W ROC area
SMO with all attributes by default parameter	**59.25**	**0.593**	**0.228**	**0.602**	**0.593**	**0.591**	**0.72**
SMO with best selected attributes by default parameter	57.29	0.573	0.238	0.581	0.573	0.572	0.702

In the first setting of SMO experiment a classifier with accuracy 59.25%, weighted precision 60.2%, and rate of weighted ROC area rate 72% was generated by taking default value with all attributes.

In the second setting of SMO experiment a classifier with accuracy 57.29%, weighted precision rate 58.1%, and weighted ROC area 70.2%

Experiment results show that SMO experiment with all attributes scored better accuracy as compared with SMO with best attribute and hence based on SMO with all attributes was selected as compared in this section.

5.2.4. Multilayer Perception (MLP) Neural Network

Multilayer Perception is a neural network based classification which uses back propagation to classify instances into known classes. There are two experiments conducted using MLP with by taking default parameter value with all attributes and 10875 records. The second experiment was carried out by taking default parameter value with best selected six attributes and 18875 records. Summary of experiment result using MLP is shown in table 5.5.

Table 5.5: experiment result using MLP Neural Network

Experiments	Accuracy	WTP rate	WFP rate	WTP precision	WTP recall	WF measure	wROC area
MLP with all attribute by default parameter	**77.239**	**0.772**	**0.126**	**0.772**	**0.772**	**0.772**	**0.875**
MLP with best selected attribute by default parameter	61.933	0.619	0.205	0.617	0.619	0.618	0.787

The above table show that, in the first setting of MLP experiment a classifier with accuracy 77.23%, weighted Precision rate 77.2% and rate of ROC area 87.5 % was generated by taking default value with all attribute for the experiment.

In the second setting of MLP experiment a classifier with accuracy 61.9%, weighted Precision rate 61.7%, and weighted recall 61.9% is constructed.

Based on the MLP experiment with all attributes in default parameter scored better accuracy as compared with MLP with best attribute by a default parameter. Taking the performance indicated above MLP with all attributes was selected as compared to the best attribute selected.

5.2.5. Naïve Bayes

Two experiments are also conducted using Naïve bayes classification algorithm. The first experiment was carried out by taking the default parameter values with all attributes and 10875 records. And the second experiment was conducted by taking the default parameter value with best five selected attributes and 18875 records. Summary of experimental result using naïve bayes is presented in table 5.6 below

Table 5.6: Experimental results using naïve bayes

Experiments	Accuracy	WTP rate	WFP rate	WTP precision	WTP recall	WF measure	wROC area
Naïve byaes with all attribute	**57.75**	0.578	0.222	**0.576**	**0.578**	**0.576**	**0.749**
Naïve byaes with best attribute	57.44	0.574	0.226	0.573	0.574	0.573	0.745

The above table show that, in the first setting of Naïve Bayes experiment a classifier with accuracy of 57.75%, weighted precision rate 57.6%, and rate of weighted ROC area74.9% was generated by taking default value and all attributes.

In the second setting of Naïve bayes experiment a classifier with accuracy 57.44 weighted precision rate 57.4%, and weighted ROC area 74.5%.

The result implies that Naïve bayes with all attributes scored better accuracy as compared with Naïve bayes with best attribute. Taking the performance indicated above classification model constructed using Naïve bayes with all attributes was selected as compared to the best attributes.

5.3. Model Evaluation

The aim of model evaluation is to confirm that the models are of high quality to achieve the objectives. The most important objectives of the current study is identification of DM model that performs best in predicting the nutritional status of women.

As it is summarized in table 5.7 model comparisons was performed using performance evaluation matrix describing true prediction rate, false prediction precision, recall and ROC area and accuracy of the models.

Table 5.7: Summary of model comparison

Experiments	Accuracy	WTP rate	WFP rate	WTP precision	WTP recall	WF measure	W ROC area
unPruned J48 with all attributes	81.917	0.819	0.098	0.818	0.819	0.818	0.908
Unpruned PART with all attributes	**83.147**	**0.831**	0.092	**0.831**	**0.831**	**0.831**	**0.907**
SMO with all attribute	59.25	0.593	0.228	0.602	0.593	0.591	0.72
MLP with all attributes	77.239	0.772	0.126	0.772	0.772	0.772	0.875
Naïve Bayes with all attributes	57.75	0.578	0.222	0.576	0.578	0.576	0.749

Summary result in table 5.7 shows complete summary of the performance of five data mining models used in current study. Although the performance of the five mining models is comparable, they gave different results in terms of accuracy.

The accuracy, True positive, precision rate, recall, and ROC characteristics of the unpruned PART rule induction with all attributes is higher than that of the other models. Therefore, the model created by this algorithm is selected as the best model that can predict the Nutritional status of women of reproductive age.

Confusion Matrix for unpruned PART rule induction

Typically, a classifier is evaluated by confusion matrix. It is useful tool for analyzing how well the selected classifier can recognize the records of different classes. Since PART classifier is the best model in this study, the confusion matrix for PART rule induction classifier is presented in table 5.8.

Table 5.8: Confusion Matrix of PART classifier with all attributes

		prediction		
		Overweight	Normal	Underweight
Actual	Overweight	4749	220	79
	Normal	386	5684	1266
	Underweight	146	1084	5261

The above table shows out of the total records 4749 were correctly classified as “Overweight”, 5648 records are correctly classified as “Normal”, and 5261 records are classified as “Underweight”. Hence 83.15% instant are correctly classified and 16.85 instance are incorrectly classified.

ROC Analysis for PART rule induction classifier

ROC analysis provides tools to select possibly optimal models and to discard submittal ones. It is related in direct and natural way to cost/benefit analysis of nutrition decision making. Figure 5.2 shows the area under ROC for the overweight instances.

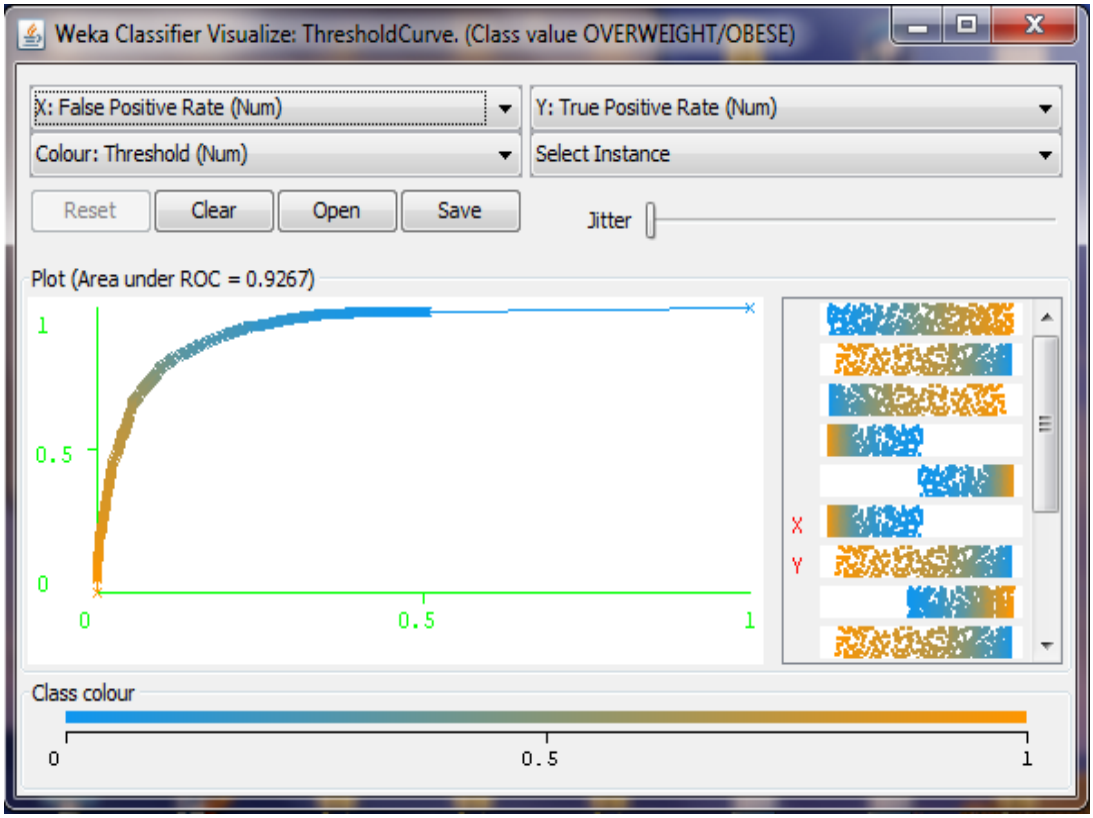


Figure 5.2: the area under ROC from the PART classifier

The area under ROC for the overweight instances produced from the PART classifier is shown in figure 5.2. The vertical axis of ROC curve represents the true positive rate. The horizontal axis represents the false positive rate.

In the above figure at first the curve moves sharply up from zero showing that there is more true positives than false positive. Then the curve starts to become more horizontal as it encounters an

increase of both true positives, and false positive. The area under the curve for this model is 0.9267.

5.4. Rules extracted from the selected Model

In this study of experimentation the selected model, unprimed PART classifier is used to generate relevant rules for the domain. From the total 14 rules were selected. These rules were generated to identify the risk of malnutrition. Based on the discussion with domain expert and show the most important attributes, the following rules were selected from the identified model generated.

Rule 1: Residence = Rural AND Current pregnant = No AND Region = Tigray AND Anemia level = Not Anemic AND Education = No Education AND Current use of contraceptive method = modern AND Current Marital status = Divorced AND AND women age group = 15-19 AND parity = 5+child then UNDERWEIGHT (28.0).

The above rule states that if divorced non pregnant women are residence Tigray not educated, use modern contraceptive method whose age group 15-19 and have at least 5 children then their nutritional status is underweight. This prediction gives 100% (28/28) correct result.

Rule 2: If Residence = Rural AND Current pregnant = No AND wealth status =Poor AND parity = 3-4child And frequency of listening radio = not at all AND sources of drinking water improved AND Region = Tigray AND Anemia level = Not Anemic AND Education = No Education AND Current use of contraceptive method = no use AND Current Marital status = Divorced then UNDERWEIGHT (20.0)

The above rule states that if divorced non pregnant women are residence Tigray wealth status is classified as poor have at least 4 children, not listening radio, not educated and not use contraceptive then their nutritional status is underweight. This prediction gives 100% (20/20) correct result.

Rule 3: If Residence = Urban AND Current pregnant = No AND Wealth status = rich AND Region = Dire Dawa AND women relationship to HH head = wife AND marital status = married AND parity = 5+child AND current breastfeeding = No OVERWEIGHT occur with (105,0)

The above rule states that if married non pregnant women are residence Dire Dawa wealth status is classified as rich and have at least 5 children and then their nutritional status is overweight. This prediction gives 100% (105/105) correct result.

Rule 4: If Residence = Rural AND Current pregnant = No AND Region = Affar AND Education = No Education AND Current Marital status = Married AND Sources of drinking water = Non improved AND Religion = MUSLIM AND Current use of contraceptive method = No use AND parity = 3-4child AND frequency of listening radio = Not at all AND wealth status = poor AND Anemia level = Not Anemic AND women age group = 25-29 AND relationship to HH = Head AND Currently breastfeeding = Yes then UNDERWEIGHT occur with (16.0/1.0)

The above rule states that if married non pregnant women are residence Affar wealth status is classified as poor, not educated, use non improved sources of drinking water, not use contraceptive method, not listening radio, breastfeed have at least 3 children and are head of HH then their nutritional status is underweight. This prediction gives 94% (16/1) correct result.

Rule 5: If Residence = Urban AND Region = Addis Ababa AND Current use of contraceptive method = No use AND Religion=Orthodox AND Wealth status =rich AND marital status=married AND anemia level =Not anemic AND education = Secondary AND frequency of listening radio = at least once a week AND sex of HH head =Male AND parity = 1-2child AND women age group = 25-29 then OVERWEIGHT occur with (78.0)

The above rule states that if married secondary level of education women are residence Addis Ababa wealth status is classified as rich have at least 1 children and age group are 25-29 then their nutritional status is Overweight. This prediction gives 100% (78/78) correct result.

Rule 6: If Residence = Rural AND Current pregnant = No AND wealth status = poor AND Region = Gambela AND education =No AND currently working =No AND Current use of contraceptive method = No use AND parity = 3-4child AND frequency of listening radio = Not a at all AND women relationship to HH head = wife AND sources of drinking water = not improved then UNDERWEIGHT occur with (39.0)

The above rule states that if married non pregnant women are residence Gambela wealth status is classified as poor, not educated, have no work, not use of contraceptive, not listening radio, use unimproved sources of water have at least 3 children and are wife of HH then their nutritional status is underweight. This prediction gives 100% (39/39) correct result.

Rule7: If AND Residence = Rural AND Current pregnant = No AND wealth status = poor AND Region = Gambela AND Marital status =married AND Religion = protestant AND relationship to HH head = Head AND parity = 5+child then UNDERWEIGHT occur with (63.0)

The above rule states that if married non pregnant women are residence Gambela wealth status is classified as poor have at least 5 children and are head of HH then their nutritional status is underweight. This prediction gives 100% (63/63) correct result.

Rule 8: If Residence = Rural AND Current pregnant = No AND wealth status = poor AND Anemia level = Not anemic AND Region = Benishangul AND Marital status = married AND toilet facility = non improved AND currently working =No AND parity = 1=2child AND uses of contraceptive method =Not use AND education = No education AND Currently breastfeeding = Yes AND Religion = Muslim AND women age group = 25-29 then UNDERWEIGHT occur with (17.0/3.0).

The above rule states that if married non pregnant women are residence Benishangul wealth status is classified as poor, use unimproved latrine, have no work, not use contraceptive method, not educated gave breastfeed 25-29 age group have at least 1 children and are head of HH then their nutritional status is underweight. This prediction gives 85% (17/3) correct result.

Rule 9: If AND Residence = Urban AND Current pregnant = No AND wealth status = rich AND Region = Harari AND Marital status =married AND Anemia level =Not anemic AND Religion = protestant AND relationship to HH head = wife AND age group =30-34 AND current use of contraceptive = modern AND education =primary child then OVERWEIGHT occur with (63.0)

The above rule states that if married non pregnant women are residence Harari wealth status is classified as rich use modern contraceptive method 30-34 age group primary level education and

are wife of HH then their nutritional status is overweight. This prediction gives 100% (63/63) correct result.

Rule 10: If Residence = Rural AND toilet facility = Non improved AND Region = Somalia AND Education = No Education AND Sources of drinking water = Non improved AND women relationship to HH head = wife AND frequency of listening radio = Not at all AND Anemia level = Not anemic AND women age group = 20-24 AND parity = 3-4child then UNDERWEIGHT occur with (45.0/3.0)

The above rule states that if married not educated women are residence Somalia use unimproved latrine and unimproved sources of drinking water , 20-24 age group have at least 3 children and are wife of HH then their nutritional status is underweight. This prediction gives 93.7% (45/3) correct result.

Rule 11: If Residence = Rural AND Current pregnant = No AND Region = Affar AND Education = No Education AND Current Marital status = Married AND Sources of drinking water = Non improved AND Religion = MUSLIM AND Current use of contraceptive method = No use AND parity = 1-2child AND frequency of listening radio = Not at all AND wealth status = poor AND Anemia level = Mild AND women age group = 25-29 AND sex of HH head = F AND Currently breastfeeding = Yes then UNDERWEIGHT occur with (22.0/1.0)

The above rule states that if married non pregnant women are residence Affar wealth status is classified as poor not educated, use unimproved sources of drinking water not use contraceptive method not listening radio, 25-29 age group gave breast feed and have at least 1 children then their nutritional status is underweight. This prediction gives 95% (22/1) correct result.

Rule12: If Residence = Rural AND Current pregnant = No AND Region = SNNP AND Anemia level = Not Anemic AND wealth index =poor AND parity =3-4child AND contraceptive method = no use AND marital status = married AND frequency of listening radio=not at all, AND toilet facility = improved AND Currently breastfeeding = Yes AND Religion = protestant AND sources of drinking water = non improved then NORMAL occur with (55.0)

The above rule states that if married non pregnant women are residence SNNP wealth status is classified as poor not use contraceptive method, not listening radio use unimproved latrine and

unimproved sources of drinking water, gave breastfeed and have at least 3 children then their nutritional status is normal. This prediction gives 100% (55/55) correct result.

Rule 13: If Residence= Urban AND Wealth status = rich AND marital status = married AND toilet facility = improved AND Region Oromiya AND sources of drinking water =improved AND age group=35-39 AND parity 1-2child then OVERWEIGHT (45.0).

The above rule states that if married 35-39 age group women are residence Oromiya wealth status is classified as rich use improved latrine and sources of drinking water then their nutritional status is overweight. This prediction gives 100% (45/45) correct result.

Rule 14: Residence = Rural AND Current pregnant = No unsure AND Region = Amhara AND women relationship to HH head = wife AND Education = No education AND Anemia level = Not anemic AND cont = Not using AND parity = 5+child AND Currently breastfeeding = Y AND frequency of listening radio = less than once a week AND women age group = 35-39 AND toilet facility = Non improved then NORMAL occur with (36.0/2.0)

The above rule states that if married non pregnant women are residence Amhara have at least 5 children not educated, not use of contraceptive, gave breast feed, listened radio at least once a week, 35-39 age group, use non improved latrine and are wife of HH then their nutritional status is Normal. This prediction gives 94.7% (36/2) correct result.

5.5. Discussion on Major Findings

The rules generated from the unpruned PART rule induction predict the nutritional status by 83.14% accuracy. The rules reflect importance of different combined attributes for classifying the nutritional status of women of reproductive age in Ethiopia.

The findings revealed that socio-economic and educations are the major risk factors for malnutrition among women of reproductive age. However other factors such as sources of drinking water, toilet facility marital status, occupation of women have also influence on malnutrition.

According to rule 1 and 2 marital status in Tigray region was risk factor women's under nutrition due to under age marriage. Other risk factors in Tigray region are educational level, wealth status, not having aware about their health and nutrition, and number of having children.

As it is presented in rule 4, 6 and 7 women in Affar and Gambela regions were under nourished due to low income, current breast feeding practice being and not being educated. There is also another factor which are not use contraceptive method, parity, unimproved sources of drinking water and there is lack of awareness about their health and nutrition.

As indicated in rule 8 risk factors of women's under nutrition in Benshagul Gumuz were Breast feeding low socio economic status, and women's low level of education.

As rule 10 un-improved toilet facility and un-improved sources of drinking water in Somalia region are the reason for women's under nutrition. The other factors are education, lack of awareness about health and nutrition and unused contraceptive in Somalia region. Burden of the family as head of Household is another factor.

Our study result shows that in Addis Ababa, economic status sources, lack of awareness and education are factors for women to be overweight. In Dire Dawa, Oromiya, and Harrari the economic status is highest factor for to be women overweight were as the study done by [20], revealed that socioeconomic status, educational level and access to improved drinking water were positively associated with being overweight in Addis Ababa. Our study has similarity with study done by [20] on socio economic status and education on predicting overweight in woman.

The findings revealed that socio-economic and educations are the major risk factors for malnutrition among women of reproductive age. Other factors including sources of drinking water, toilet facility marital status, occupation of women have also influence on malnutrition. In a study made by [22] socio economic and demographic variables have significant influences in women's malnutrition. Region of residence, household economic status, and women's employment status have significant influence on the chronic energy deficiency in women. Women's age and marital status are also very important determinants of nutritional status of women of reproductive age.

In agreement with the current study, previous Ethiopian study [23] reported strong associations of lack of latrine facilities and age were with stunting and thinness. found lack of latrine facilities was significantly with stunting and thinness.

Findings from the analysis of rural sample in the current study indicated region, residence, education, lack of awareness, age, marital status, occupation and sources of drinking water, number of children, socioeconomic status, and toilet facility are predictor of women malnutrition. In urban socio economic status, education and women occupation are predictor of women malnutrition.

5.6. Prototype development

The last objective of this study was developing a prototype interface that provides easy access to the identified knowledge base. The final selected if-then rules are used to implement the selected best models from each of the five experimentations.

The development of graphical user interface (GUI) was done by using Microsoft visual studio 2010. This prototype GUI is developed based on the model generated by PART rule induction classifier with unpruned parameter and all attributes. The rules used in this study to design the GUI for predicting nutrition status of women are from 14 rules listed above.

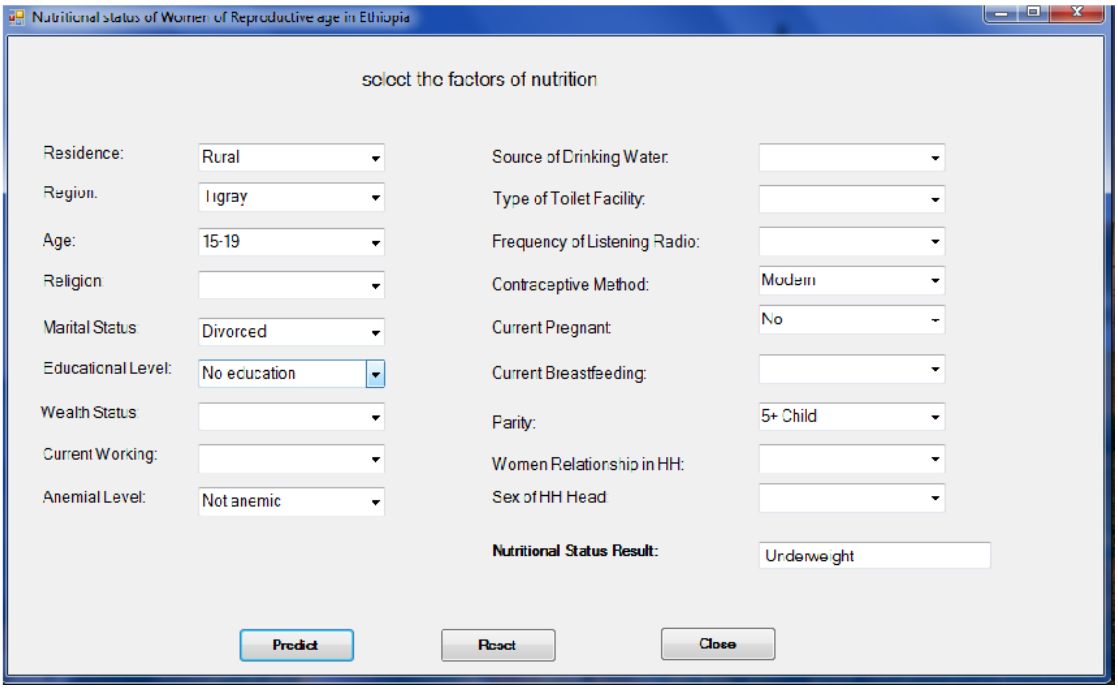


Figure: 5.3: Nutritional status predictions

As shown in Figure 5.4, this prototype prediction model can be used for predicting nutritional status of women based on the rules generated by PART classifier. The sample result shows the prototype predicts the nutrition status of women as underweight.

5.6. Evaluation of user interface

In order to know the nutritional status of women by using BMI result this study developed the user interface prototype. The questionnaire was developed based on the objective to check the validity of the interface. The questionnaire has of five questions with five options. Namely, Excellent, very good, good, satisfactory, and fair. The evaluation is conducted kazanchis and bole health center, the questionnaire filled by 6 health professionals. the summary of the evaluation result is shown table 5.9.

Table 5.9: evaluation on women nutritional status user interface

Variables	**Number**	**percent**
The feature of the interface is easy use		
Excellent	5	83.3
Very good	1	16.7
The feature of the interface is easy to understand		
Excellent	4	66.7
Very good	2	33.3
The interface in terms of time taking		
Excellent	4	66.7
Very good	2	33.3
The importance for nutrition result		
Excellent	4	66
Very good	2	33.3
The interface provide novel suggestion for		
Excellent	3	50.0
Very good	2	33.3
Good	1	16.7

The above table shows the result of evaluation question indicated all respondents approach to the validity of the interface is excellent and very good one respondent said the interface provide knowledge is good. I discussed with the health professionals after they filled questionnaire.

Two respondents have the same idea, they explained their opinion, the interface is excellent we have not seen before this type of tool for this purpose and this prototype is very useful for our task. The other two respondents explained their idea they said prototype need for us because we can predict easy to the nutritional status of the women. The other respondents from center explained he impressed with the functionality of the interface. The idea is very good, but he said for my opinion it is not that much necessary for me. I personally suggest further discussion is need for the implementation of the interface with health professional will use the prototype.

CHAPTER SIX

Conclusion and Recommendation

6.1. Conclusion

The main goal of this study was to apply data mining techniques to predict nutritional status of women of reproductive age. The data used in this study was extracted from the country wide 2011 Ethiopia Demographic and Health Survey (EDHS). We used J48 decision tree, PART rule induction SMO Support Vector Machine, MLP Artificial Neural Network, and Naïve bayes to classify the dataset. Prediction was made by classifying attributes of women in to one of the possible class.

Current age, socioeconomic status, educational level sources of drinking water, latrine facility, breast feeding status, occupation, contraceptive method being under use, marital status, anemia level, residence and region, are the determinant factors of women's malnutrition.

The risk of under nutrition is significant in rural areas of Affar, Tigray, Benishangul Gumuz, Gambela and Somalia. In these regions poor socioeconomic status, unimproved sources of drinking water, unimproved latrine facility, parity unused contraceptive for family planning, unawareness of women of their own health and nutritional status could be the reason associated with under nutrition. Addis Ababa, Dire Dawa, urban area of Oromiya, and Harari, the socioeconomic status is a main risk factor of women's overweight. In these areas, educated women and women who have reported of having access to improve sources of drinking water were at an increased factor risk of being overweight.

The unpruned PART rule induction classifier was selected as the best model to be able to predict the nutritional status of women in Ethiopia. PART rule induction generated 2500 rules that predict nutritional status of women.

The findings described above entails that data mining is useful in bringing relevant information from large and complex dataset so that anybody can use this information for decision making.

6.2. Recommendation

Based on the findings of the current study it is possible to forward the following recommendations:

- ❖ This study was conducted to find out the potential applicability of data mining technology to predict the nutritional status of women of reproductive age. There is however, a need to the development of knowledge based system for nutrition status prediction with domain experts. This should be a future research direction.

- ❖ This study, attempts has been made to find out the potential applicability of data mining technology to predict the nutritional status of women of reproductive age. There is a need for the development of knowledge based system for women's nutrition status with domain experts. This should be a further research direction

- ❖ To enhance the performance of the present model, further study should be conducted using nutrition data of women of reproductive age using many more mining techniques to improve the predictive model accuracy.

Reference

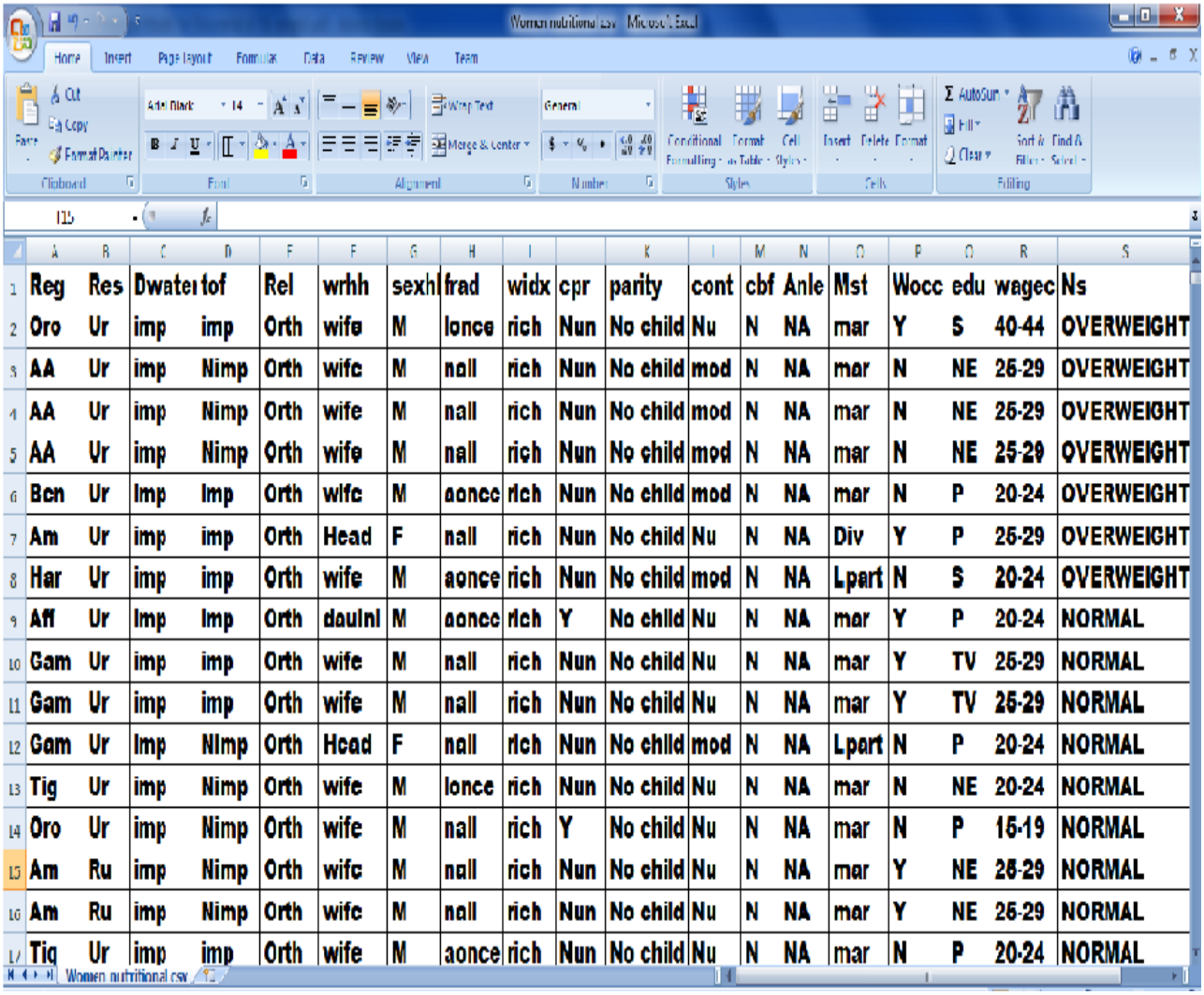
1. An assessment of the causes of malnutrition in Ethiopia, A contribution to the formation of National Nutrition Strategy for Ethiopia, Edited by Todd Bensom, International Food policy Research institute, Washington, DC, USA, Nov, 2005.
2. Ethiopia National Nutrition Strategy, Review and analysis of Progress and Gaps, save the children UK, May, 2009
3. WHO, Essential Nutrition Action; improving maternal, newborn, infant and young child health and nutrition. May 2011.
4. Government of the Federal Democratic Republic of Ethiopia, National Nutrition Programme. June 2013-June 2015
5. David J., Heikki Mannila, and Padhraic Smyth Principles of data mining, A Bradford Book The MIT press, Cambridge, Massachusetts London England, 2001.
6. Ian H.Witten. Eibe Frank. Mark A. Hall, DM practical machine learning tools and techniques, third edition, Morgan Kaufmann publishers, 2011
7. Faqdzilah siraj and Mansour Ali Abdoulha, Mining Enrolment Data using predictive and descriptive approaches, Jan 21, 2011, DOI: 10.5772/14210
8. Ethiopia Demographic and Health Survey Addis Ababa, Ethiopia 2011, Central Statistical Agency, Addis Ababa, Ethiopia ICF international, Calverton, Maryland, USA, March, 2012
9. Oded Maimon, Lior Rokach, Data Mining and Knowledge Discovery, Second Edition, Springer New York Dordrecht Heidelberg London, April, 2010. ISBN-13: 978-0387098227 ISBN-10: 0387098224
10. Julio Ponce and Adem Karahoca, Data Mining and Knowledge Discovery in Real Life Application, Jan 1, 2009. ISBN 978-3-902613-53-0, 436 pages, Publisher: I-Tech Education and Publishing, DOI: 10.5772/97
11. Jiban k pal, “Usefulness and application of data mining in extracting information from different perspectives.” March 2011, Vol. 58, PP 7-16
12. David L. Olson and Dursun Delen Advanced Data Mining Techniques Publisher: Springer; 2008 edition (March 4, 2008), ISBN-10**:** 3540769161, ISBN-13: 978-3540769163

13. ZHAI Lianga, TANG Xinmingb, WU Land, LI Lina, WANG Zhongyuana: Data mining and its application in database content refinement, 2000
14. Ralf Mikut and Makus Reischi, Data Mining tools, January/February, 2011 Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz,1, 76344 Eggenstein Leopoldshafen, Germany, DOI: 10.1002/widm.24
15. Centers for Disease Control and Prevention (CDC), Open source Data Mining software Evaluation, 2012
16. Hian Chye Koh and Gerald Tan Data Mining application in Health Care ,2010 Journal of Healthcare Information Management — Vol. 19, No. 2
17. Hian Chye Koh and Gerakd Tan, Data Mining applications in Healthcare. Journal of Healthcare Information Management, 2011 — Vol. 19, No. 2
18. Zenebe Marcos, Prediction Under nutritional status of under five children using data mining techniques: In the case of 2011 Ethiopian Demographic and Health Survey, Health med Informant, 1000152, ISSN: 2157-7420 JHMI, an open access journal, volume 5. Issue 2, 2014.
19. D.Thangamani, Identification of malnutrition with use of supervised data mining techniques decision trees and artificial neural networks, International journal of Engineering and computer science ISSN-2319-7242,vol-3 issue-9 September 9, 2014 page No 8236-8241
20. Tebekaw Yibeltal, "The burden of underweight and overweight among women in Addis Ababa, Ethiopia." BMC Public Health 2014, 14:1126, doi:10.1186/1471-2458-14-1126
21. Wasie Belaynew, "Nutritional status of Adults living with HIV/AIDS at the university of Gondar referral hospital, Northwest Ethiopia" Ethiop. J. Health Biomed Sci., 2010. Vol.3, No.1
22. Wodemariam Girma, Timotiows Genebo Determinants of nutritional status of women and children in Ethiopia, Ethiopian health and nutrition research institute, Addis Ababa, Ethiopia, ORC Macro, Calaverton, Maryland, USA November 2002
23. Afework Mulugeta, Nutritional status of Adolescent Girls from Rural Communities of Tigray, Northern Ethiopia, Ethiopian Journal of Health development 23(2009) 1,- ISSN 1021-6790 -page 5-11

24. Charu C. Aggarawal, Data Classification algorithms and application, July 25, 2014 by Chapman and Hall/CRC , ISBN 9781466586741 - CAT# K20307 , Reference - 707 Pages
25. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, second Edition, Morgan Kaufmann Publisher, 500 Sansome Street, Suite 400, San Francisco, CA 94111, 2011
26. John Mingers, An empirical comparison of pruning Methods for Decision Tree Induction, vol.38 No.3, page.309-338, March *2000*,
27. Minneapolis Department of Health and Family Support, Risks of being overweight for women of reproductive age, January, 2008
28. Dr. Rajesh Verma, Classification algorithms for Data Mining; A Survey, Vol. 1 Issue 2 August 2012 International Journal of Innovations in Engineering and Technology (IJIET) ISSN: 2319 – 1058
29. Megha Gupta, Naveen Aggarwal, Classification techniques analysis, March 20, 2010
30. Kimito Funatsu and Kiyoshi Hasengawa, New fundamental technologies in data mining, Published by InTech 2011, ISBN 978-953-307-547-1
31. David L. Olson Dursun Delen, Advanced Data Mining Techniques. 2008 , ISBN: 978-3-540-76916-3
32. Carlos Gershhenson, Artificial Neural Networks for Beginners , Nov 17, 2006
33. Centers for Disease control and prevention (CDC), Body Mass Index: considerations for practitioners. cdc 24/7: solving lives protecting people Feb 8, 2011,
34. Body Mass index internet available from http://www.halls.md/body-mass-index/bmi.htm
35. Two crows, Introduction to data mining and knowledge discovery, third edition, 1999, ISBN: 1-892095-02-5
36. David J. Statistics and Data Mining: Intersecting Disciplines, Department of Mathematics, Imperial College, London, UK, ACM SIGKDD, 1999.
37. Jiban k pal, “Usefulness and application of data mining in extracting information from different perspectives.” March 2011, vol58, issue 1, page 7-16, publisher NISCAIR, CSIR, india
38. Tipawan Silwattananusarn and Assoc.Prof. Dr. KulthidaTuamsuk, “Data Mining and Its Applications for Knowledge Management.” A Literature Review from 2007 to 2012

ANNEX I

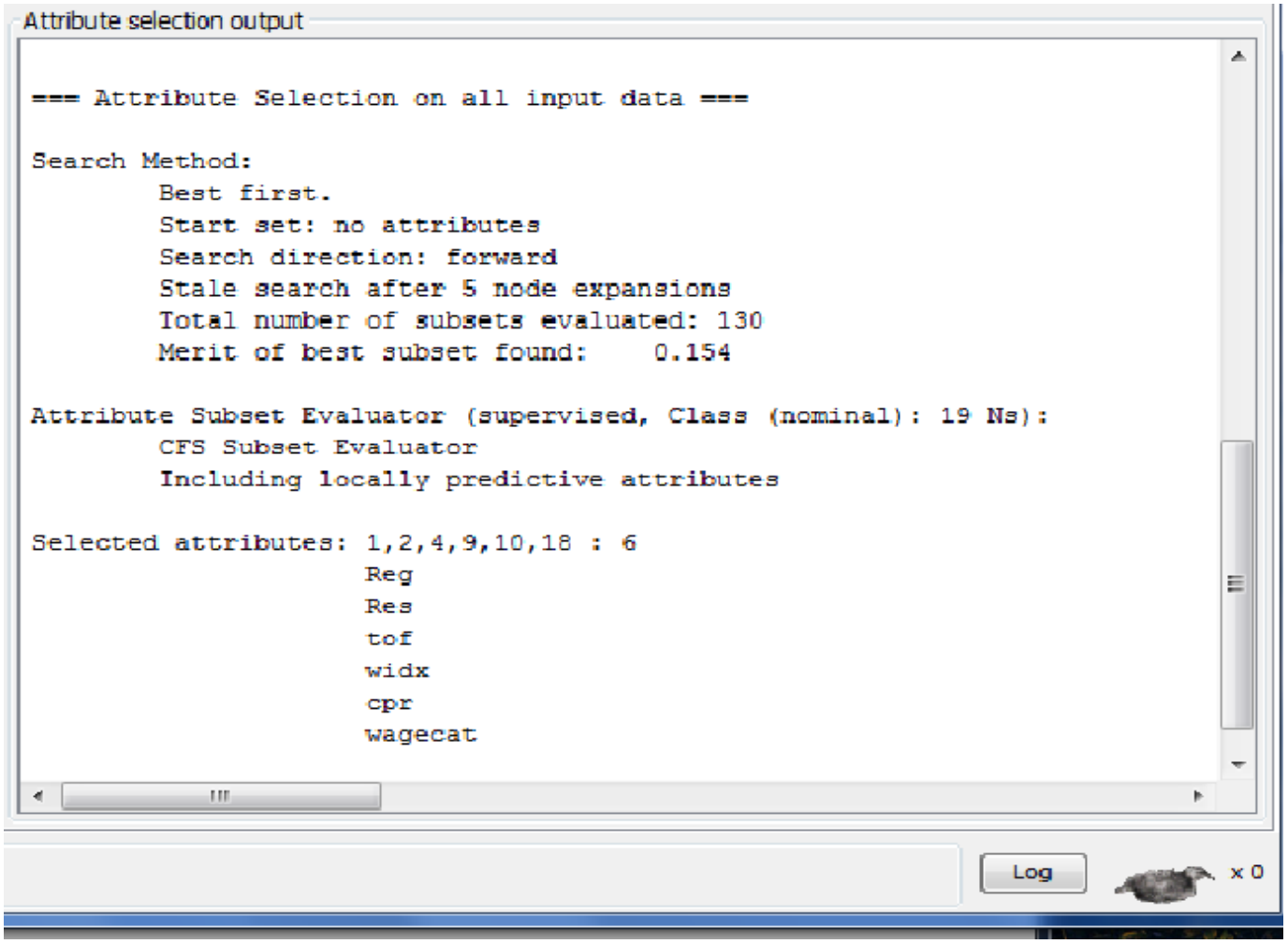
Dataset sample with CSV (comma delimited) File Format



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Reg	Res	Dwater	tof	Rel	wrhh	sexh	frad	widx	cpr	parity	cont	cbf	Anle	Mst	Wocc	edu	wagec	Ns
2	Oro	Ur	imp	imp	Orth	wife	M	lonce	rich	Nun	No child	Nu	N	NA	mar	Y	S	40-44	OVERWEIGHT
3	AA	Ur	imp	Nimp	Orth	wife	M	nall	rich	Nun	No child	mod	N	NA	mar	N	NE	25-29	OVERWEIGHT
4	AA	Ur	imp	Nimp	Orth	wife	M	nall	rich	Nun	No child	mod	N	NA	mar	N	NE	25-29	OVERWEIGHT
5	AA	Ur	imp	Nimp	Orth	wife	M	nall	rich	Nun	No child	mod	N	NA	mar	N	NE	25-29	OVERWEIGHT
6	Ben	Ur	Imp	Imp	Orth	wife	M	aonce	rich	Nun	No child	mod	N	NA	mar	N	P	20-24	OVERWEIGHT
7	Am	Ur	imp	imp	Orth	Head	F	nall	rich	Nun	No child	Nu	N	NA	Div	Y	P	25-29	OVERWEIGHT
8	Har	Ur	imp	imp	Orth	wife	M	aonce	rich	Nun	No child	mod	N	NA	Lpart	N	S	20-24	OVERWEIGHT
9	Aff	Ur	Imp	Imp	Orth	daulnl	M	aonce	rich	Y	No child	Nu	N	NA	mar	Y	P	20-24	NORMAL
10	Gam	Ur	imp	imp	Orth	wife	M	nall	rich	Nun	No child	Nu	N	NA	mar	Y	TV	25-29	NORMAL
11	Gam	Ur	imp	imp	Orth	wife	M	nall	rich	Nun	No child	Nu	N	NA	mar	Y	TV	25-29	NORMAL
12	Gam	Ur	Imp	Nimp	Orth	Head	F	nall	rich	Nun	No child	mod	N	NA	Lpart	N	P	20-24	NORMAL
13	Tig	Ur	imp	Nimp	Orth	wife	M	lonce	rich	Nun	No child	Nu	N	NA	mar	N	NE	20-24	NORMAL
14	Oro	Ur	imp	Nimp	Orth	wife	M	nall	rich	Y	No child	Nu	N	NA	mar	N	P	15-19	NORMAL
15	Am	Ru	imp	Nimp	Orth	wife	M	nall	rich	Nun	No child	Nu	N	NA	mar	Y	NE	25-29	NORMAL
16	Am	Ru	imp	Nimp	Orth	wife	M	nall	rich	Nun	No child	Nu	N	NA	mar	Y	NE	25-29	NORMAL
17	Tig	Ur	imp	imp	Orth	wife	M	aonce	rich	Nun	No child	Nu	N	NA	mar	N	P	20-24	NORMAL

ANNEX II

Result of CFS Attributes Subset Evaluator



ANNEX III

Sample Weka output

=== Run information ===

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

Relation: Women nutritional status-weka.filters.supervised.instance.SMOTE-C0-K5-P60.0-S1-weka.filters.supervised.instance.SMOTE-C0-K5-P65.0-S1-weka.filters.supervised.instance.SMOTE-C0-K5-P70.0-S1-weka.filters.supervised.instance.SMOTE-C0-K5-P75.0-S1

Instances: 18875

Attributes: 19

Region, Residence, Sources of drinking water, toilet facility, Religion, women Relationship to HH, Sex of HH, Frequency of listening radio, wealth index, current pregnancy, parity, use of contraceptive, current breast feeding, anemia level, marital status, women occupation, education, women age category, Nutritional status

Test mode:10-fold cross-validation

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

=== Stratified cross-validation ===

Mean absolute error	0.1208
Root mean squared error	0.3052
Relative absolute error	27.5004 %
Root relative squared error	65.1212 %
Total Number of Instances	18875

=== Summary ===

Correctly Classified Instances	15694	83.147 %
Incorrectly Classified Instances	3181	16.853 %
Kappa statistic	0.745	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F measure	Roc Area	Class
	0.941	0.038	0.899	0.941	0.92	0.966	Overweight/Obese
	0.775	0.113	0.813	0.775	0.794	0.877	Normal
	0.811	0.109	0.796	0.811	0.803	0.894	Underweight
Weighted Avg.	0.831	0.092	0.831	0.831	0.831	0.907	

=== Confusion Matrix ===

a	b	c	<-- classified as
4749	220	79	a = OVERWEIGHT/OBESE
386	5684	1266	b = NORMAL
146	1084	5261	c = UNDERWEIGT

ANNEX IV

Visual Basic code

```
Private Sub Button1_Click(ByVal sender As System.Object, ByVal e As System.EventArgs)
Handles predictbclick.Click, predictbclick.MouseCaptureChanged,
predictbclick.MarginChanged
        If (rescbox.SelectedItem = "Rural" And
            cpcbox.SelectedItem = "No" And
            regcbox.SelectedItem = "Tigray" And
            anacbox.SelectedItem = "Not anemic" And
            educbox.SelectedItem = "No education" And
            cmcbox.SelectedItem = "Modern" And
            mcbox.SelectedItem = "Divorced" And
            agecbox.SelectedItem = "15-19" And
            pcbox.SelectedItem = "5+ Child") Then
            nsrtbox.Text = " Underweight"
        ElseIf (rescbox.SelectedItem = "Rural" And
                cpcbox.SelectedItem = "No" And
                wealthcbox.SelectedItem = "Poor" And
                pcbox.SelectedItem = "3-4 Child" And
                flrcbox.SelectedItem = "Not at all" And
                sdwcbox.SelectedItem = "Improved" And
                regcbox.SelectedItem = "Tigray" And
                anacbox.SelectedItem = "Not anemic" And
                educbox.SelectedItem = "No education" And
                cmcbox.SelectedItem = "No Use" And
                mcbox.SelectedItem = "Divorced") Then
            nsrtbox.Text = " Underweight"
        ElseIf (rescbox.SelectedItem = "Urban" And
                cpcbox.SelectedItem = "No" And
                wealthcbox.SelectedItem = "Rich" And
               regcbox.SelectedItem = "Dire Dawa" And
               wrhhcbox.SelectedItem = "Wife" And
                 mcbox.SelectedItem = "Married" And
                 pcbox.SelectedItem = "5+ Child" And
             cbfcbox.SelectedItem = "No") Then
            nsrtbox.Text = " Overweight"
        ElseIf (rescbox.SelectedItem = "Rural" And
            cpcbox.SelectedItem = "No" And
            regcbox.SelectedItem = "Affar" And
            educbox.SelectedItem = "No education" And
            mcbox.SelectedItem = "Married" And
            sdwcbox.SelectedItem = "Non-improved" And
            relcbox.SelectedItem = "Muslim" And
            cmcbox.SelectedItem = "No Use" And
            pcbox.SelectedItem = "3-4 Child" And
            flrcbox.SelectedItem = "Not at all" And
            wealthcbox.SelectedItem = "Poor" And
            anacbox.SelectedItem = "Not anemic" And
            agecbox.SelectedItem = "25-29" And
            wrhhcbox.SelectedItem = "Head" And
            cbfcbox.SelectedItem = "Yes") Then

            nsrtbox.Text = " Underweight"
        ElseIf (rescbox.SelectedItem = "Urban" And
                regcbox.SelectedItem = "Addis Ababa" And
```

```
        cmcbox.SelectedItem = "No Use" And
        relcbox.SelectedItem = "Orthodox" And
        wealthcbox.SelectedItem = "Rich" And
        mcbox.SelectedItem = "Married" And
        anacbox.SelectedItem = "Not anemic" And
        educbox.SelectedItem = "Secondary" And
        flrcbox.SelectedItem = "At least once a week" And
        shhhcbox.SelectedItem = "Male" And
        pcbox.SelectedItem = "1-2 Child" And
        agecbox.SelectedItem = "25-29") Then
    nsrtbox.Text = " Overweight"
ElseIf (rescbox.SelectedItem = "Rural" And
        cpcbox.SelectedItem = "No" And
        wealthcbox.SelectedItem = "Poor" And
        regcbox.SelectedItem = "Gambela" And
        educbox.SelectedItem = "No education" And
        cwcbox.SelectedItem = "No" And
        cmcbox.SelectedItem = "No Use" And
        pcbox.SelectedItem = "3-4 Child" And
       flrcbox.SelectedItem = "Not at all" And
       wrhhcbox.SelectedItem = "Wife" And
       sdwcbox.SelectedItem = "Non-improved"
       ) Then
    nsrtbox.Text = " Underweight"
ElseIf (rescbox.SelectedItem = "Rural" And
        cpcbox.SelectedItem = "No" And
        wealthcbox.SelectedItem = "Poor" And
        regcbox.SelectedItem = "Gambela" And
        mcbox.SelectedItem = "Married" And
        relcbox.SelectedItem = "Protestant" And
        wrhhcbox.SelectedItem = "Head" And
        pcbox.SelectedItem = "5+ Child") Then
    nsrtbox.Text = " Underweight"
ElseIf (rescbox.SelectedItem = "Rural" And
        cpcbox.SelectedItem = "No" And
        wealthcbox.SelectedItem = "Poor" And
        anacbox.SelectedItem = "Not anemic" And
        regcbox.SelectedItem = "Benishangul-Gumuz" And
        mcbox.SelectedItem = "Married" And
        ttfcbox.SelectedItem = "Non-improved" And
        cwcbox.SelectedItem = "No" And
        pcbox.SelectedItem = "1-2 Child" And
        cmcbox.SelectedItem = "No Use" And
        educbox.SelectedItem = "No education" And
        cbfcbox.SelectedItem = "Yes" And
        relcbox.SelectedItem = "Muslim" And
        agecbox.SelectedItem = "25-29") Then

    nsrtbox.Text = " Underweight"
ElseIf (rescbox.SelectedItem = "Urban" And
   cpcbox.SelectedItem = "No" And
   wealthcbox.SelectedItem = "Rich" And
   regcbox.SelectedItem = "Harari" And
   mcbox.SelectedItem = "Married" And
   anacbox.SelectedItem = "Not anemic" And
   relcbox.SelectedItem = "Protestant" And
    wrhhcbox.SelectedItem = "Wife" And
    agecbox.SelectedItem = "30-34" And
```

```
            cmcbox.SelectedItem = "Modern" And
            educbox.SelectedItem = "Primary") Then

            nsrtbox.Text = " Overweight"
        ElseIf (rescbox.SelectedItem = "Rural" And
            ttfcbox.SelectedItem = "Non-improved" And
            regcbox.SelectedItem = "Somalia" And
            educbox.SelectedItem = "No education" And
            sdwcbox.SelectedItem = "Non-improved" And
            wrhhcbox.SelectedItem = "Wife" And
            flrcbox.SelectedItem = "Not at all" And
            anacbox.SelectedItem = "Not anemic" And
            agecbox.SelectedItem = "20-24" And
            pcbox.SelectedItem = "3-4 Child") Then
            nsrtbox.Text = " Underweight"

        ElseIf (rescbox.SelectedItem = "Rural" And
            cpcbox.SelectedItem = "No" And
            regcbox.SelectedItem = "Affar" And
           educbox.SelectedItem = "No education" And
         mcbox.SelectedItem = "Married" And
        sdwcbox.SelectedItem = "Non-improved" And
        relcbox.SelectedItem = "Muslim" And
        cmcbox.SelectedItem = "No Use" And
        pcbox.SelectedItem = "1-2 Child" And
        flrcbox.SelectedItem = "Not at all" And
        wealthcbox.SelectedItem = "Poor" And
        anacbox.SelectedItem = "Mild" And
        agecbox.SelectedItem = "25-29" And
      shhhcbox.SelectedItem = "Female" And
      cbfcbox.SelectedItem = "Yes") Then

            nsrtbox.Text = " Underweight"
        ElseIf (rescbox.SelectedItem = "Rural" And
        cpcbox.SelectedItem = "No" And
        regcbox.SelectedItem = "SNNP" And
        anacbox.SelectedItem = "Not anemic" And
        wealthcbox.SelectedItem = "Poor" And
        pcbox.SelectedItem = "3-4 Child" And
        cmcbox.SelectedItem = "No Use" And
         mcbox.SelectedItem = "Married" And
         flrcbox.SelectedItem = "Not at all" And
         ttfcbox.SelectedItem="Improved" And
         cbfcbox.SelectedItem="Yes" And
         relcbox.SelectedItem = "Protestant" And
         sdwcbox.SelectedItem="Non-improved") Then

            nsrtbox.Text = " Normal"
        ElseIf (rescbox.SelectedItem = "Urban" And
      regcbox.SelectedItem = "Oromiya" And
   wealthcbox.SelectedItem = "Rich" And
   mcbox.SelectedItem = "Married" And
   ttfcbox.SelectedItem = "Improved" And
   sdwcbox.SelectedItem = "Improved" And
    agecbox.SelectedItem = "35-39" And
    pcbox.SelectedItem = "1-2 Child") Then

            nsrtbox.Text = " Overweight"
```

```
        ElseIf (rescbox.SelectedItem = "Rural" And
                cpcbox.SelectedItem = "No" And
                regcbox.SelectedItem = "Amhara" And
                wrhhcbox.SelectedItem = "Wife" And
                educbox.SelectedItem = "No education" And
                anacbox.SelectedItem = "Not anemic" And
                cmcbox.SelectedItem = "No Use" And
                pcbox.SelectedItem = "5+ Child" And
                cbfcbox.SelectedItem = "Yes" And
                flrcbox.SelectedItem = "Less than once a week" And
                agecbox.SelectedItem = "35-39" And
                ttfcbox.SelectedItem = "Non-improved") Then

            nsrtbox.Text = " Normal"

        Else
            nsrtbox.Text = "Unknown"
        End If

    End Sub

    Private Sub rescbox_SelectedIndexChanged(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles rescbox.SelectedIndexChanged

    End Sub

    Private Sub cpcbox_SelectedIndexChanged(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles cpcbox.SelectedIndexChanged

    End Sub

    Private Sub predictbclick_MouseDoubleClick(ByVal sender As Object, ByVal e As
System.Windows.Forms.MouseEventArgs) Handles predictbclick.MouseDoubleClick

    End Sub

    Private Sub Button2_Click(ByVal sender As System.Object, ByVal e As System.EventArgs)
Handles Button2.Click
        rescbox.SelectedIndex = "-1"
        regcbox.SelectedIndex = "-1"
        agecbox.SelectedIndex = "-1"
        relcbox.SelectedIndex = "-1"
        mcbox.SelectedIndex = "-1"
        educbox.SelectedIndex = "-1"
        wealthcbox.SelectedIndex = "-1"
        cwcbox.SelectedIndex = "-1"
        anacbox.SelectedIndex = "-1"
        sdwcbox.SelectedIndex = "-1"
        ttfcbox.SelectedIndex = "-1"
        flrcbox.SelectedIndex = "-1"
        cmcbox.SelectedIndex = "-1"
        cpcbox.SelectedIndex = "-1"
        cbfcbox.SelectedIndex = "-1"
        pcbox.SelectedIndex = "-1"
        wrhhcbox.SelectedIndex = "-1"
        shhhcbox.SelectedIndex = "-1"
        nsrtbox.Text = " "
    End Sub
```

ANNEX V

Evaluation Questionnaire

Addis Ababa University
School of Graduate Studies
School of Information Science and School of Public Health
Prototype Validity Evaluation Questionnaire

Dear Participants,

This questionnaire is designed to check the validity of the prototype which designed as part of MS.c study. The truthfulness of your responses will contribute much to the validity of the prototype. You are, therefore, cordially requested to be honest to provide accurate information. I would like to let you know that any information you provide in this questionnaire will be kept strictly confidential and will only be used for this study.

Gender: 1. Male 2. Female

Questions	Poor	Fair	Good	Very Good	Excellent
The features of interface in terms of easy to use					
The features of interface in terms of easy to understand					
The features of interface in terms of time taking					
The importance for nutrition result					
The interface provide novel suggestion for prediction					

Thank You!