ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND

SCHOOL OF PUBLIC HEALTH

PREDICTING UNDER NUTRITION STATUS OF UNDER-FIVE CHILDREN USING DATA MINING TECHNIQUES: THE CASE OF 2011 ETHIOPIAN DEMOGRAPHIC AND HEALTH SURVEY

ZENEBE MARKOS

JUNE 2013

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

PREDICTING UNDER NUTRITION STATUS OF UNDER-FIVE CHILDREN USING DATA MINING TECHNIQUES: THE CASE OF 2011 ETHIOPIAN DEMOGRAPHIC AND HEALTH SURVEY

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN HEALTH INFORMATICS

BY
ZENEBE MARKOS

JUNE 2013

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

PREDICTING UNDER NUTRITION STATUS OF UNDER-FIVE CHILDREN USING DATA MINING TECHNIQUES: THE CASE OF 2011 ETHIOPIAN DEMOGRAPHIC AND HEALTH SURVEY

BY

ZENEBE MARKOS

Members of the Examining Board:

Name	Title	Signature	Date
_____	Chairperson	_____	_____
Dr. Martha Yifiru	Advisor,	_____	_____
Dr. Jemal Haidar	Advisor,	_____	_____
Dr. Solomon Teferra	Examiner,	_____	_____
Dr. Solomon Shiferaw	Examiner,	_____	_____

Declaration

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

_____

Zenebe Markos

June, 2013

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

_____ _____

Martha Yifiru (PhD) Jemal Haidar (MD)

June, 2013 June, 2013

ACKNOWLEDGEMENTS

First of all, I gratefully express my deepest thanks to the almighty God for his guidance, help, support and because He let me see new days of success in my life. Glory to God.

I am heartily thankful to my advisors, Dr. Martha Yifiru and Dr. Jemal Haidar for their encouragement, guidance, constructive comments, support and their help that enabled me to develop and understanding of the subject.

I forward my sincere gratitude to Central Statistics Agency staff for providing the dataset used in this study and Ethiopia Health and Nutrition Research Institute staffs, especially Aweke Kebede (PhD fellow), for their general comments on the nature of the dataset.

I would like to thank Addis Ababa University, School of Information Science and School of Public Health for financial support and overall facilitation of the research from the beginning until the end.

I also thank to Minale Tefera, Temesgen Dileba, Tesfahun H/Mariam, Wondimu H/Mariam, Temesgen Tamirat, Beemnet Moges, Abebe Lolamo, Deneke Lefebo including my classmates Misge, Dav and Mamush for sharing ideas in the course of writing this thesis and their comments on my work.

Last, but not least, I would like to express my heartfelt gratitude to all my family members specially to my lovely friend, S/r Helen Haile (Baxxiyiye) and Zemede Markos who supported me in any respect during my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AREF	Attribute Relationship File Format
BMI	Body Mass Index
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSAE	Central statistics Agency of Ethiopia
CSV	Comma Separated Value
EDHS	Ethiopia Demographic and Health survey
EHNRI	Ethiopia Health and Nutrition Research Institute
FMOH	Federal Ministry of Health
GUI	Graphic User Interface
HAZ	Height for Age Z-score
IDD	Iodine Deficiency Disorders
KDD	Knowledge Discovery in Database
KDP	Knowledge Discovery Process
PEM	Protein-Energy Malnutrition
KNN	K-Nearest Neighbors
ROC	Receiver Operating Characteristics
SEMMA	Sample, Explore, Modify, Model, and Assess
SMOTE	Synthetic Minority Over-sampling TEchnique
SPSS	Statistical Package for Social Sciences
SNNPRG	South Nations Nationalities People Regional Government
WAZ	Weight for Age Z-score
WEKA	Waikato Environmental for Knowledge Analysis
WHO	World Health Organization
WHZ	Weight for Height Z-score
WHOCGMS	World Health Organization Child Growth Multicenter Standards

ABSTRACT

Background: under nutrition is one of the leading causes of morbidity and mortality in children under the age of five in most developing countries including Ethiopia.

Objective: The general objective of this study was to design a model that predicts the nutritional status of under-five children using data mining techniques.

Methodology: This study followed hybrid methodology of Knowledge Discovery Process to achieve the goal of building predictive model using data mining techniques and used secondary data from 2011 Ethiopia Demographic and Health Survey dataset. Hybrid process model was selected since it combines best features of Cross-Industry Standard Process for Data Mining and Knowledge Discovery in Database methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives. WEKA 3.6.8 data mining tools and techniques such as J48 decision tree, Naïve Bayes and PART rule induction classifiers were utilized as means to address the research problem.

Result: In this particular study, the predictive model developed using PART pruned rule induction found to be best performing having 92.6% of accurate results and 97.8% WROC area. Promising result has been achieved from the rules regarding nutritional status prediction.

Conclusion: The results from this study were encouraging and confirmed that applying data mining techniques could indeed support a predictive model building task that predicts nutritional status of under-five children in Ethiopia. In the future, integrating large demographic and health survey dataset and clinical dataset, employing other classification algorithms, tools and techniques could yield better results.

Keywords*: Predictive modeling, Nutritional status, children, Data mining, EDHS dataset*

CHAPTER ONE

INTRODUCTION

1.1 Background

Good nutrition is an essential component of good health. Nutrition is at the heart of most global health problems – especially in the area of child survival where child under nutrition is an underlying cause of more than one-third (3.5 million) prevalence of all child deaths under the age of five in developing countries. Of the 112 million underweight children and 178 million children who suffer from stunting, 160 million (90%) live in just 36 developing countries, constituting almost half (46%) of the cases (1).

Malnutrition is a known contributing factor to disease and death. Worldwide, over 10 million children under the age of five years die every year from preventable and treatable illnesses despite effective health interventions (2). One in four of the world's children are stunted (3). At least half of these deaths are caused by malnutrition. In addition, malnourished children that survive are likely to suffer from frequent illness, which adversely affects their nutritional status and locks them into a vicious cycle of recurring sickness, faltering growth and diminished learning ability (4).

Children malnutrition is a major public health problem in developing countries (5). It contributes to child morbidity and mortality, poor intellectual and physical development of children, lowered resistance to diseases, and consequently stifles development (6).

Nutritional status is an outcome and impact indicator when assessing progress towards achieving the Millennium Development Goals (MDGs). Marked differences, especially with regard to height-for-age and weight-for-age, are often seen among different subgroups within a country. Child nutritional status is related with MDGs special Goal 4 of 2/3 child mortality reduction in 2015. Improving child nutrition is a key to achieving this goal. Child mortality is deeply interlocked with all the other MDGs (1). Each one is a major contributor to poor and dangerous living conditions for children.

Federal Ministry of Health (FMOH) and Central Statistics Agency of Ethiopia (CSAE) in collaboration with non-governmental organizations collected large volume of dataset to identify children nutritional factors and reported nutritional status of children using SPSS for analysis with limited tools (7). These limited attributes can be improved using a technology that has the capacity to extract hidden useful information from such type of data through data mining. Data mining is a crucial step in discovery of knowledge from large datasets. In recent years, data mining has found its significant hold in every field including health care. Mining process is more than the data analysis which includes classification, clustering, and association rule discovery. It also spans other disciplines like data warehousing, statistics, machine learning and artificial intelligence.

Predicting the outcome of a diagnosis is one of the most interesting and challenging tasks to develop data mining tasks. In recent years new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets. For prediction of nutritional status of under-five children, a data mining techniques, such as J48 decision tree, Naïve Bayes and PART rule induction were applied.

1.2 Statement of the Problem and its Justification

In Ethiopia, malnutrition remains one of the leading cause of morbidity and mortality of under-five children, particularly under nutrition, which is widespread in the country and contributes to over half (53%) of all children deaths in the country. Ethiopia has the second highest rate of malnutrition in Sub-Saharan Africa (8). The causes of malnutrition probably vary in different settings, as well as over time. However, little information is available on nutritional status of the demographic and socioeconomic segments of the country to formulate targeted tackling solutions and overcome under-five malnutrition. Useful information of the relative contribution of the major demographic risk factors of under-nutrition is therefore an important for developing nutrition intervention strategies in Ethiopia.

Although several studies in Ethiopia were conducted to map out the magnitude of the major nutritional problems, almost all of them were limited in their study area/scope. Most of the studies (4, 6, 8, 9, 10, 11) were cross-sectional descriptive research design, logistic regression and multivariate analysis. All of the studies on child nutrition were descriptive in nature and limited to the analysis of associations between nutritional status with certain nutrition-related variables and risk factors of malnutrition in children. To overcome the aforementioned problems, data mining hybrid methodology that combines the best features of Cross-Industry Standard Process for Data mining (CRISP-DM) and KDD was applied. The main advantage of using data mining techniques over the statistical methods used in previous researches is that they enable to predict malnutrition level than only identifying risk factors for a specific outcome at group level. Typically logistic regression used in the above studies helped to identify risk factors for nutritional status, thereby making easy to identify victims at high risk. But, models developed with the use of data mining techniques were used to predict nutrition status of under-five children. The techniques and algorithms were J48 decision tree, Naïve Bayes and PART rule induction.

Previous studies' data size and number of variables were limited. This was because the cost, time and capabilities of the software that they applied for analysis. Child malnutrition has long been recognized as one of the most serious problems in Ethiopia; however national-level data on levels and determinants of malnutrition is scarce.

It is therefore reasonable to study that what attribute values are affecting nutritional status and develop a model that assists in predicting future interventions based on the values of significant attributes identified. The ability of the tools and algorithms of data mining to deal with datasets characterized by thousands of instances and high dimensionality (large number of attributes) coupled with the understandability of models produced at the end and their ease of use would make data mining suitable for this study. Now a days, data mining technology is being used as a tool that provides the techniques to transform these mounds of data into useful information which in turn enables to derive knowledge for decision making. In addition to this, it is better to study using large datasets and attributes with advanced technology called data mining.

The main purpose of this study was to apply data mining techniques for extracting hidden patterns which are significant to predict nutritional status of under-five children from 2011 EDHS dataset. To achieve this goal, the following research questions were formulated for investigation.

- What are the optimal determinant factors that lead to child under nutrition in Ethiopia?
- Which predictive modeling algorithms are suitable for determining nutritional status of under-five children?

1.3 Objective of the Study

1.3.1General Objective

The general objective of this study was to explore 2011 EDHS dataset to construct a model that predicts the under nutrition status of under-five children using data mining techniques.

1.3.2 Specific Objectives

The specific objectives of this study were to:

- Conduct a thorough review of literature on data mining techniques and children under nutrition that can be used to attain the objective of nutritional status prediction
- Prepare good quality dataset by applying preprocessing tasks such as data cleaning, data transformation and attribute selection
- Build predictive models that can be applicable on a new instance in order to determine the nutritional status of under-five children

- Compare the performance of models based on classifiers' evaluating criteria in predicting nutritional status of under-five children
- Report research findings and make recommendations.

1.4 Methodology of the Study

1.4.1 Research Design/Data Mining Modeling

The aim of this study was to uncover hidden patterns in the 2011 EDHS dataset. The study followed hybrid methodology of Knowledge Discovery Process (KDP) to achieve the goal of building predictive model using data mining techniques. Hybrid process model was selected since it combines best features of CRISP-DM and KDD methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives. Hybrid methodology basically involves six steps: problem domain understanding, data understanding, data preparation, data mining, evaluation and use of the discovered knowledge. The six steps of hybrid KDP model (12) is depicted in Figure 1.1.

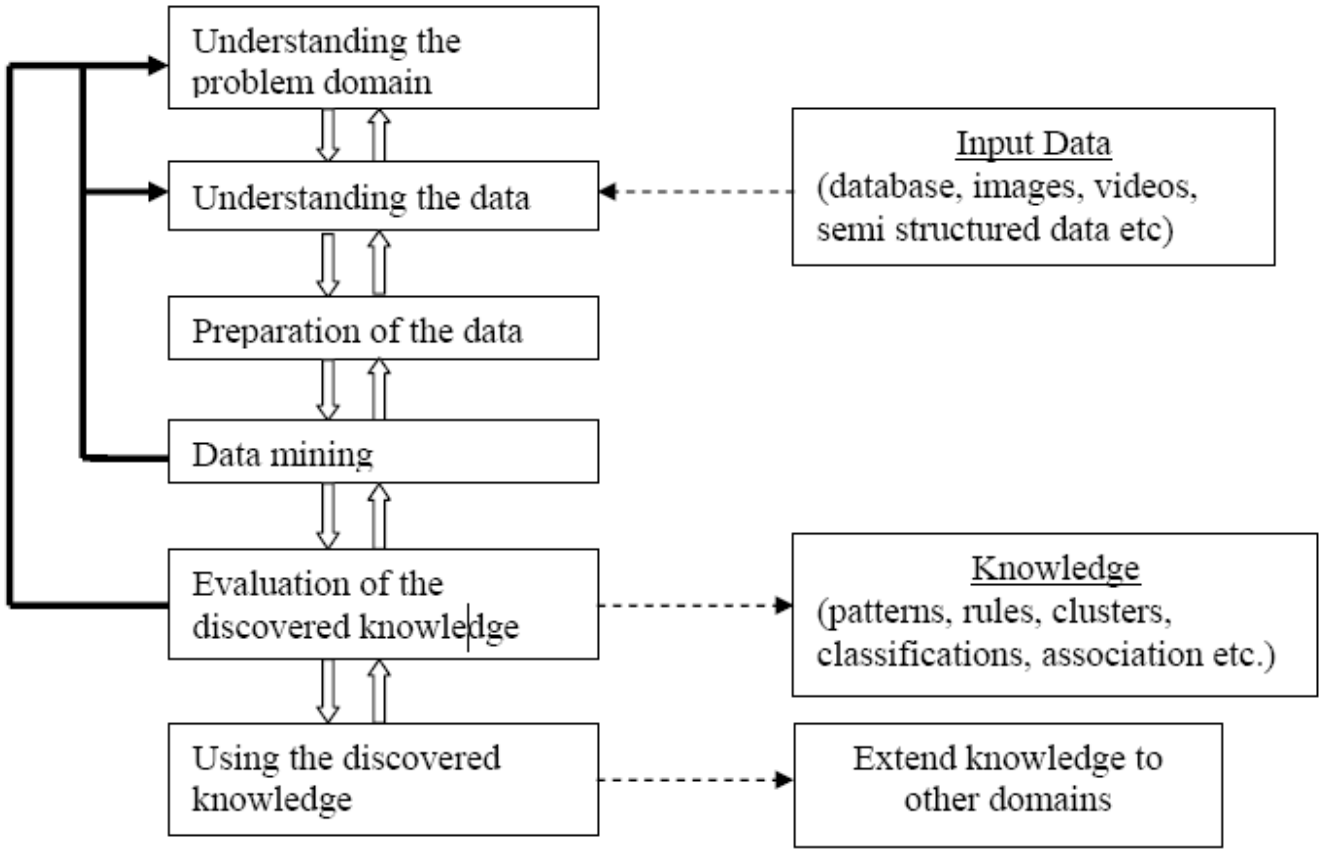

Figure 1.1: The six steps of hybrid KDP model

One of the important aspects of this model is its iterative and interactive feature. The feedback loops are necessary because any changes and decisions made in one of the steps can result in changes in subsequent steps. There are several activities and processes which can be performed in each of these steps. The following is the description of the activities performed in each step.

1.4.1.1 Understanding Problem Domain /Children Malnutrition

In this step of hybrid model, discussions with domain experts from CSA and Ethiopia Health and Nutrition Research Institute (ENHRI) were conducted as the main source and reviewing different documents, books and journal articles that focus on data mining techniques in health care as well as nutritional status of children was also used as supporting sources.

1.4.1.2 Understanding Data

This step starts with an initial data collection and proceeds with activities in order to get familiar with and detect appropriate subsets of the 2011 EDHS dataset. In this study, the 2011 EDHS dataset was used as a source of data. FMOH and ECSA in collaboration with non-governmental organizations collected data related to nutritional status of under-five children, fertility and family planning behavior, child mortality, adult mortality, maternal mortality, utilization of maternal and child health services, knowledge of HIV/AIDS and prevalence of HIV/AIDS and prevalence of anemia in children and mothers at national level. The survey was conducted in nine administrative regions (Affar, Amhara, Benishangul-Gumuz, Gambela, Harari, Oromiya, SNNPRG, and Somali) and two city administrations (Addis Ababa and Dire Dawa).

One part of the questionnaire is about under-five children and maternal health. Information related to a child such as age, sex, region, residence, breastfeeding duration, hemoglobin level, amenorric, height for age SD, weight for height SD, etc are stored in the database. Maternal attributes in the questionnaire such as age, education, wealth index, ethnicity, occupation, exposure, place of delivery, etc are found in the dataset. The 2011 EDHS dataset contains a total of 920 attributes and 11,654 instances (7)

1.4.1.3 Data Preparation

All raw datasets which are initially prepared for data mining are often large; many are related to humans and have the potential for being messy. Data preparation step, is concerned with deciding the data that will be used as input for data mining modeling or methods in the subsequent steps. To this end, data cleaning (such as filling missing values, detecting outliers) and relevance analysis were done using different descriptive statistics such as data visualization using SPSS and WEKA tools.

After the data is cleaned, it was transformed into WEKA understand format by exploring the 2011 EDHS dataset (which is in SPSS) into the Comma Separated Value (.csv) and Attrubute-Relation File Format (.arff). The end result would be maintaining a good quality data that meets the specific input requirements for the selected data mining tools. The cleaned data was further processed by feature selection and extraction algorithms.

1.4.1.4 Data Mining

One of the basic activities in hybrid KDP model is data mining where algorithms and techniques are used for searching interesting patterns and creating predictive model. In this study J48 decision tree algorithm, Naïve Bayes classifier and PART rule induction have been used.

In data mining, J48 decision tree is a predictive model which can be used to represent both classification and regression models. Decision trees are also useful for exploring data and gaining insight into the relationships of a large number of candidate input variable to the target variable. Decision trees can represent rules usually expressed in the form of "if condition then outcome," which constitute the text version of the model that can readily be expressed and easily understood by humans.

Bayesian classifier is statistical classifier and a practical learning algorithm that can predict class membership probabilities. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes and classification is based on a probabilistic model specification. The advantage of Naïve Bayes classifier is that it reaches the minimum error when the dataset is large and the methods for estimating particular probabilities are consistent.

PART rule induction algorithm extracts rules. Due to this reason, the algorithm is categorized under classification by rule induction. The algorithm builds partial decision trees and reads a path from the root of the tree to the leaf to construct a rule. The rules are linked together to give a complete set of rules. PART has almost a similar set of parameters with J48 algorithm that can be adjusted to build better model from datasets.

This step is a step where J48 decision tree, Naïve Bayes and PART rule induction were applied to create models. Different experiments were performed on the imbalanced dataset by SMOTE analysis in WEKA 3.6.8 machine learning software using 10-fold, percentage split test models with their default and adjusted parameter. Best first, information gain ratio and chi-square attribute evaluation methods were also adopted to examine if there would be any improvement in the performance of the algorithms.

1.4.1.5 Evaluation of the Discovered Knowledge

After mining the required patterns/rules, the interpretation and evaluation of the mined patterns was accomplished. The interpretation was concerned with whether the detected pattern was interesting or not. It was verified whether it has knowledge or not. The performance of the algorithms was measured and evaluated using accuracy such as, sensitivity, specificity, TP Rate, FP Rate, Recall, F-Measure and Receiver Operating Characteristic (ROC) area. Rules were generated from the best model. Detecting interesting rules and interpreting them were carried out.

1.4.1.6 Using the Discovered Knowledge

Finally, discovered knowledge will be disseminated by using different techniques. Therefore, the overall research design was developed models that predict the nutritional status of under-five children (probability of normal, wasted underweight or stunted).

1.4.2 Dissemination of the Result

The result of this study would be presented and disseminated to different concerned organization/bodies such as, School of Information Science and School of Public Health by

submitting original copy. Effort will be made to disseminate the results of the study through the following ways:

- Presentation for the School of Public Health and School of Information Science;
- Putting the hardcopy in the libraries of School of Information Science and School of Public Health so that interested readers can get access to the research output;
- Presentation on different conferences/workshops;
- Publishing in different journals.

1.5 Scope and limitations of the Study

The scope of this research is limited to develop a model that can assist in predicting nutritional status of under-five children in Ethiopia using the 2011 EDHS dataset and data mining techniques. The inclusion criterion of this study was records of children whose age group is under five years. Moreover, the study is limited to the development of predictive model using hybrid methodology due to time limitation.

Another limitation of this study was lack of literatures related to the application of data mining techniques on prediction of nutritional status of under-five children and the application of data mining techniques in demographic and health survey datasets.

1.6 Significance of the Study

The findings from this research might be used to assist in predicting the nutritional status of under-five children. Primarily, the research work has an explicit significance in development of the knowledge for the researcher. Moreover it can be used as a benchmark for interested researchers to explore the issues in the area. Information on nutritional status of children and causes of malnutrition is important to inform regional and national health policy makers and other stakeholders who are collaboratively working on child health to monitor the impact of interventions and progress towards MDGs. The outcome of this study would provide hidden knowledge by extracting large volumes of data and a model which can be used to predict the determinants of malnutrition of children.

1.7 Organization of the Thesis

The research work is organized in to six chapters. The first chapter deals with background of study through introducing the burden of under-five malnutrition conditions, statement of the problem and its justification, objective of the study, methodology of the study, scope and limitations of the study and significance of the study.

The second chapter discusses briefly about overview of data mining and knowledge discovery in database, methodologies of data mining research, data mining tasks, methods of data mining tool selection and application of data mining in health care.

The third chapter mainly focuses on under-five children malnutrition in relation to burden of developed and developing countries. Specifically, overview of malnutrition, causes of malnutrition, nutritional status of children, measurements of nutritional status of under-five children and related works were discussed in this chapter.

Chapter four attempted the first two steps of hybrid model (problem domain understanding and data understating) to address the driving force of under-five malnutrition through understanding the business area. It also shows data preparation, data preprocessing for mining, data transformation, description of preprocessed and prepared data, tasks done to generate a good quality dataset.

Chapter five presents the experimentation and evaluation of discovered knowledge, experimental design, selecting and evaluating the attributes, algorithm classifier parameters, model building rule extraction and error rate of the selected model.

At the end, chapter six provides concluding remarks and recommendations.

CHAPTER TWO

DATA MINING

2.1 Overview of Data Mining and Knowledge Discovery in Database

In the recent years the ever increasing accumulation of raw data in every industry has created both an opportunity and a challenge to the process of knowledge discovery. The challenge associated with the largeness of the data size is related to the limited processing capabilities of prevailing statistical tools which lead to a demand for better methods to deal with the large volume of data. But, challenges are not the only thing that large data bases have come up with, opportunities are also associated with them i.e. they possess patterns and hidden information that represent interesting and hidden knowledge. Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database. The discipline concerned with this task has become known as *data mining* (13).

Data mining defined by different scholars and it has several definitions for data mining, but the one which is most frequently used by scientific community is that: *Data mining is an activity that extracts new nontrivial information contained in large databases in order to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine learning, statistics and database technologies (14).*

Moreover, it is considered as a synonym for knowledge discovery in databases (KDD) by many people and it was defined as an automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams (15). On the other hand, data mining is viewed by others as single but an essential step in the larger process of KDD. It is the data mining step which is concerned with the application of specific algorithms for extracting patterns from the data (16).

One of the aims of data mining is the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful.

This relationships and summaries derived through data mining are often referred to as models or patterns. A model is a high-level description, summarizing a large collection of data and describing its important features. The structure of the model is a global summary of a dataset. In contrast to the global nature of models, local patterns make statements only about restricted regions of the space spanned by the variables (17).

The patterns discovered during data mining must be meaningful in that they usually lead to an economic benefit. Researchers often strive to discover the patterns that govern how the physical world works and encapsulate them in theories that can be used for predicting what will happen in new situations (18).

2.2 Methodologies of Data Mining Research

Data mining is a dynamic research and development area reaching maturity. As a research, it requires stable and well defined foundations which are well understood and popularized throughout the community (19).The primary goal of data mining methodologies is building stable models following some logical steps (13). Hence, different methodologies of data mining research attempt to shape the activities the researcher performs in a typical data mining process (20). Popular methodologies applied in data mining research include KDD, SEMMA (Sample, Explore, Modify, Model, and Assess) and Cross-Industry Standard Process for Data Mining (CRISP-DM) (21).

2.2.1 Knowledge Discovery in Databases

Data mining is often set in the broader context of KDD. The KDD process basically involves selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures (17). Fayyad et.al (22) broadly outlined the basic steps of KDD as depicted in Figure 2.1.

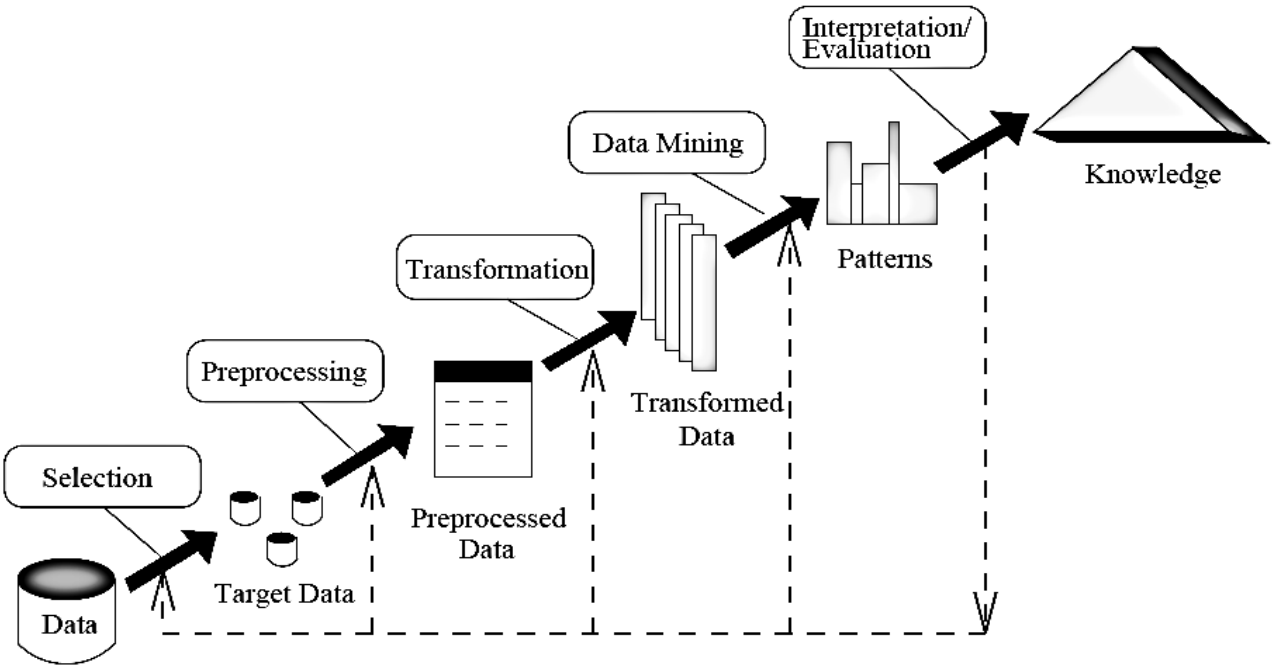

Figure 2.1: KDD process (22)

Scholars argued that the KDD process is an interactive and iterative, involving numerous steps with many decisions being made by the user. Fayyad et.al also defined the KDD process as it is preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It must be continued by the knowledge consolidation, incorporating this knowledge into the system. KDD has been more formally defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (22, 23, 12).

Each step of KDD process proposed by Fayyad e.al (22) described as follows:

Step 1: Data Selection. Given data, the first step in KDD is data selection. In this stage creating a target dataset on focus of a subset of variables needed on which discovery aimed to solve the problem are selected. For discovery purposes, data relevant to the analysis task are retrieved from the database and unnecessary data attributes should be removed.

Step 2: Data Processing. In order to produce effective data mining models in terms of quality and performance, the raw data need to undergo preprocessing in the form of data cleaning. Because real world data are mostly dirty and unclean which need to correct bad data that encountered from data redundancy, incompleteness or missing attributes value, noise, and inconsistency in order to make knowledge searching paths easey for mining algorithms.

Therefore, data quality needs to be assured in this step before going to the next phase of knowledge discovery process.

Step 3: Data Transformation. During transformation phase, data are consolidated into forms appropriate for mining to reduce data size by dividing the range of data attribute into intervals each containing approximately same number of samples or to scale attribute data to fall within a specified range. Therefore, values of attributes are changed to a new set of replacement values to ease data mining. Because of the use of different sources, data that are fine on its own may become problematic when we want to integrate it. In this step, data need to be combined from multiple sources, such as database, data warehouse, files and non-electronic sources into a coherent store. We need to merge different sourced data by keeping uniform format before running data mining tools and techniques.

Step 4: Data Mining. Data mining is the next essential process where intelligent methods are applied in order to extract hidden patterns in the data. This phase requires analysis of the main problem for patterns of interest in the data depending on the business objectives and data mining requirements. Different data mining algorithms and techniques are used for searching knowledge or interesting patterns to construct predictive or descriptive models.

Model creation is followed by performance evaluation which measures the accuracy rate of the system. The mined pattern enables to identify the truly interesting ones. For any errors or mismatched result generation as compared to domain area perspectives, the process restarts to initial step so as to provide accurate results.

Step 5: Knowledge Presentation. Finally, visualization and knowledge representation are used to present the mined knowledge to the users and stored as new knowledge in the knowledge base. Incorporating the knowledge into another system for implementation purpose, documentation and report for presenting the benefit of the knowledge to interested parties, incorporating the knowledge with previously known knowledge in the area are some of the important activities during this phase. But only this step of KDD process is not applied on this study due to hybrid process selection.

2.2.2. Cross-Industry Standard Process for Data Mining Model

The CRISP-DM model is one of the most widely used data mining methodology for knowledge discovery that consists of six steps as depicted in Figure 2.2 which are summarized below:

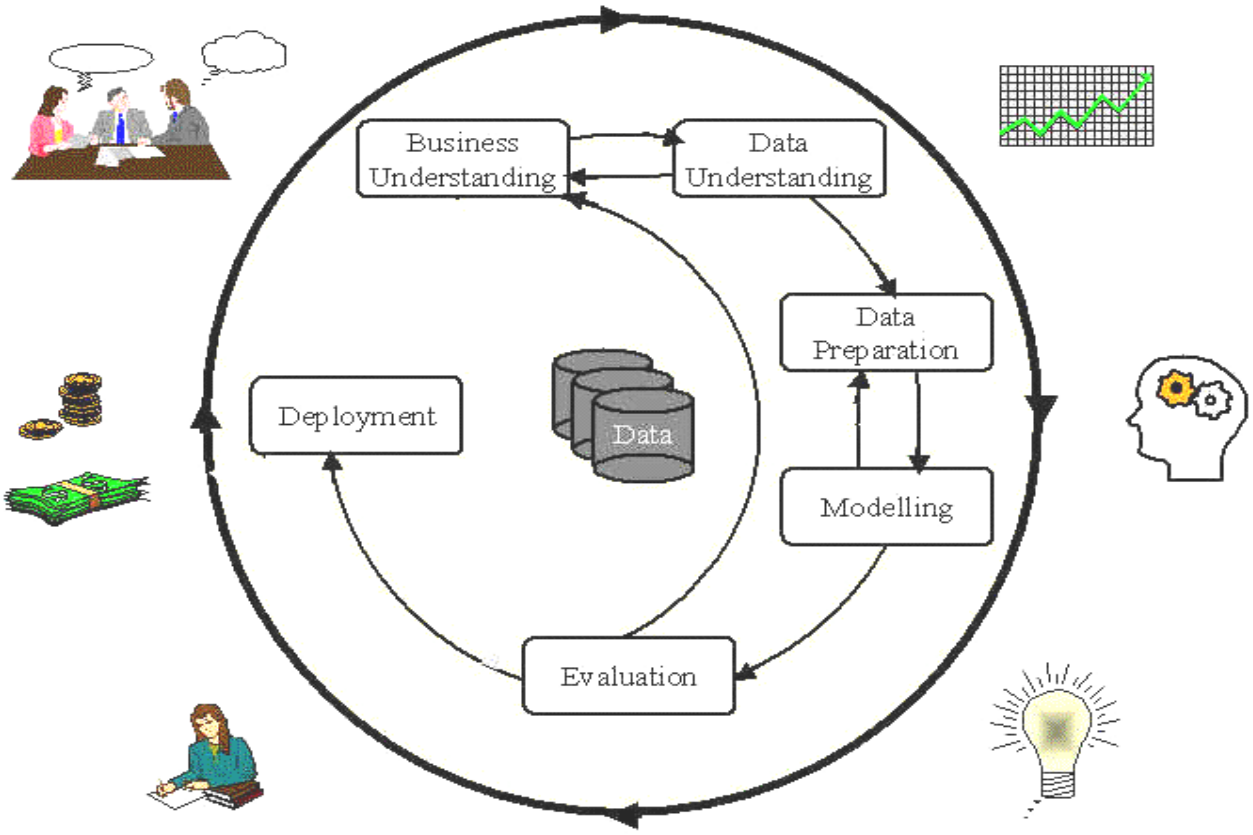

Figure 2.2 The CRISP-DM KDP model (12)

Each step discussed as follows according to Cios *et.al* (12):

Step 1: Business understanding. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a data mining problem definition, and designs a preliminary project plan to achieve the objectives. It is further broken into several sub steps, namely, determination of business objectives, assessment of the situation, determination of data mining goals, and generation of a project plan. This step is not applied on this research due to methodology specification.

Step 2: Data understanding. This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is further broken down into collection of initial data, description of data, exploration of data, and verification of data quality.

Step 3: Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into data mining tool(s) in the next step. It includes table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into selection of data, cleansing of data, construction of data, integration of data, and formatting of data sub steps.

Step 4: Modeling. At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same data mining problem type and the calibration of their parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into selection of modeling technique(s), generation of test design, creation of models, and assessment of the generated models.

Step 5: Evaluation. After one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the data mining results should be reached. The key sub steps in this step include evaluation of the results, process review, and determination of the next step.

Step 6: Deployment. Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating a report or as complex as implementing a repeatable KDP. This step is further divided into plan deployment, plan monitoring and maintenance, generation of final report, and review of the process sub steps. The researcher has not applied this step in the research due to methodology selection.

2.2.3 Sample, Explore, Modify, Model and Assess (SEMMA) Model

SEMMA process model (12) focuses on the model development aspects of data mining and involves the following logical steps as depicted in Figure 2.3:

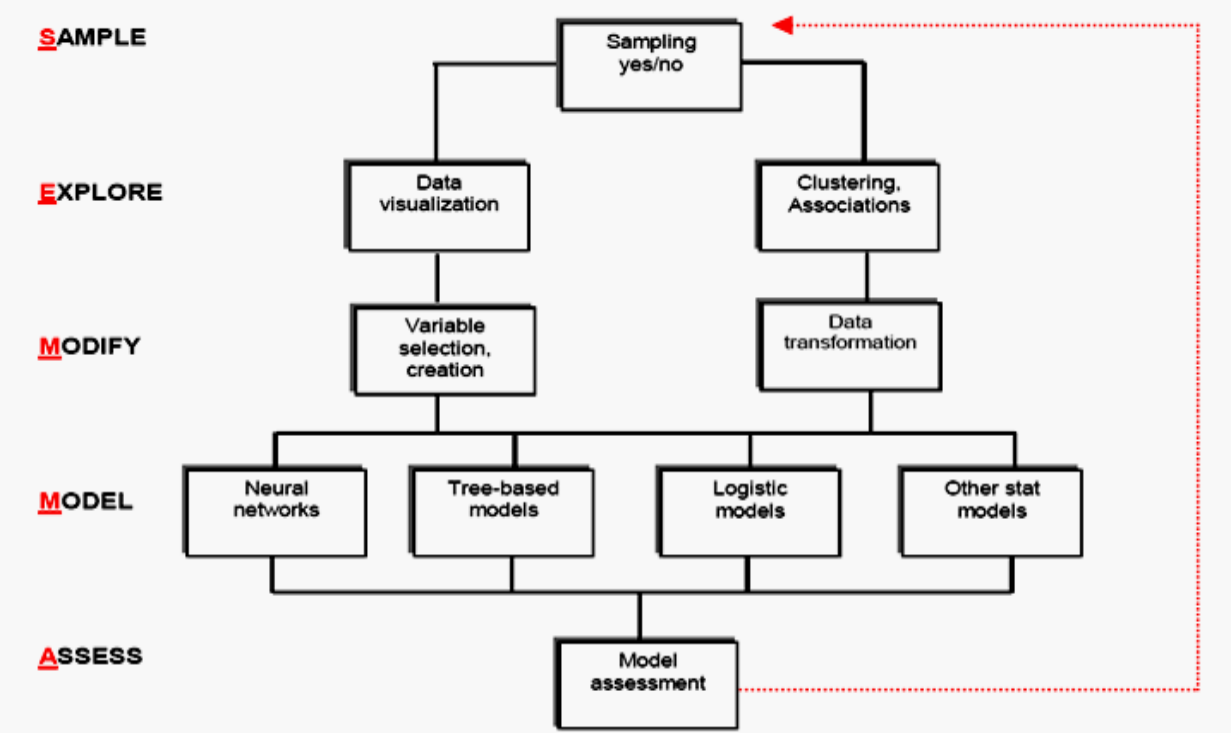

Figure 2.3: SEMMA Process model (12)

SAS (24) describe each step as follows:

Step 1: Sampling: Extracting a portion of large dataset such that the sample taken is big enough to contain the significant information, or else small enough to manipulate quickly.

Step 2: Exploring: Searching for unanticipated trends and anomalies in the data in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration does not reveal clear trends, it is possible to explore the data through statistical techniques. For example, box plots to identify outliers.

Step 3: Modifying: Creating, selecting, and transforming the variables to focus the model selection process. Based on the discoveries in the exploration phase, one may need to manipulate the data to include information such as the grouping of clients and significant subgroups, or to introduce new variables. One may also need to look for outliers and reduce the number of variables, to narrow them down to the most significant ones. One may also need to modify data when the mined data change. Because data mining is a dynamic, iterative process, one can update data mining methods or models when new information is available.

Step 4: Modeling: Allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Step 5: Assessing: Evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, one can test the model against known data.

By assessing the outcome of each stage in the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. In SEMMA, the sample steps goes equivalently with selection step of KDD and continues till to last assessment phase as interpretation/evaluation of the discovered knowledge in KDD. However, KDD manifests the pre-KDD and Post KDD that SEMMA does not. Here, SEMMA process presented for the methodology comparison and hybrid process methodology applied in this research.

2.2.4 Hybrid model

The development of academic models such as the nine-step model and eight-step model and industrial models such as five-step model and the six-step CRISP-DM model has led to the development of hybrid model that combines aspects usable for data mining research. It was developed by Cios et al. (12) based on the CRISP-DM model.

Hybrid process is characterized by providing more general, research oriented description of the steps (12). The hybrid model also encourages the application of knowledge discovered for a particular domain in other domains and it has a six step (i.e, understanding the problem domain, understanding the data, preparation of the data, data mining, evaluation of the discovered knowledge and using the discovered knowledge) process and it was presented in chapter one in Figure 1.1. Details description of each step of the hybrid model also discussed in chapter one.

Summary of correspondences between KDD, SEMMA, CRISP-DM, and Hybrid models are presented in Table 2.1 (25).

Table 2.1: Summary of data mining models

KDD	SEMMA	CRISP-DM	Hybrid
Pre KDD	----------	Business understanding	Problem domain Understanding
Selection	Sample	Data Understanding	Data understanding
Preprocessing	Explore		
Transformation	Modify	Data preparation	Data Preparation
Data mining	Model	Modeling	Data mining
Interpretation/evaluation of the discovered knowledge	Assessment	Evaluation	Evaluation
Post KDD	----------	Deployment of discovered knowledge	Use of discovered knowledge

As can be seen from Table 2.1, some of the data mining models follow the same steps in the data mining process while others follow different steps.

2.3 Data Mining Tasks

A model is a representation of the real world. Without a perfect and error free model or representation of the real world, well-intended decisions may lead to disastrous results. It forms the cornerstones of data mining. From a non-technical perspective, a data mining model may be considered as a black box, whose input is data to be mined and whose output is the knowledge discovered from the data. Data mining tasks are used to specify the kind of patterns to be found in data. Data mining tasks can be classified into two categories: predictive and descriptive modeling (26).

2.3.1 Predictive Modeling

Predictive modeling focus on building a model that will permit the value of one attribute to be predicted from the known values of other attributes. It was observed that these methods could make use of two types of techniques on the bases of the type of values the designated attribute will assume. The first of these techniques used in predictive methods is classification which is

appropriate when designated attribute is categorical. Numerical prediction (often called regression) is another method in which a model is built to predict a numeric value (27).

In predictive modeling task, one identifies patterns found in the data to predict future values. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results (28).

Supervised learning is a technique for creating a function for a training data. The training data consists of pairs of input objects (a vector of characteristics) and desired output. The output of the function can be a continuous value (regression) or can predict a class label (classification). The task of the learner is to predict the value of the outcome for any valid input object after having seen a number of training examples. In a global model, the goal is to estimate a function *g*, given a set of points (x*; g*(x)) (29).

The basic problem of supervised learning deals with predicting the response variables from the independent variables. When the output is quantitative, the problem is known as regression. The categorical variable output will lead us to classification and separation. In essence, the input *X* is a collection of *p* associated variables, and for each *X*, an observed value *Y*, of the output, is the supervisor. The goal is to build a learner, guided by the training set based on *N* samples of the pair (*Y;X*), so that it can predict the value $y\hat{}(x)$ from a future observation *x*. (29)

Tasks in predictive modeling are classification, prediction, time series analysis and regression. Only two tasks selected by the researcher based on different benefits described.

2.3.1.1. Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification methods aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Classification techniques create classification models by examining already classified instances and inductively finding a predictive pattern (28). In classification or numerical prediction model building process, a given dataset is divided into training and test sets. First, the training set is analyzed by a classification/numeric prediction

algorithms and the classifier or learner model is built. Then, test set is used in estimating the accuracy of the model built. Finally, the learner model is represented in the form of classification rules, decision trees or mathematical formulae together with various performance measures showing its ability to correctly classify new instances (12, 30). There are five main different types of classification such as Decision Tree, Naïve Bayes, K-Nearest Neighbor and Support Vector Machine. Decision Tree and Naïve Bayes classification techniques selected by the researcher due to algorithms specifications to the study and based on different benefits are described here under.

2.3.1.1.1 Decision Tree

A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) represents a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node (30).

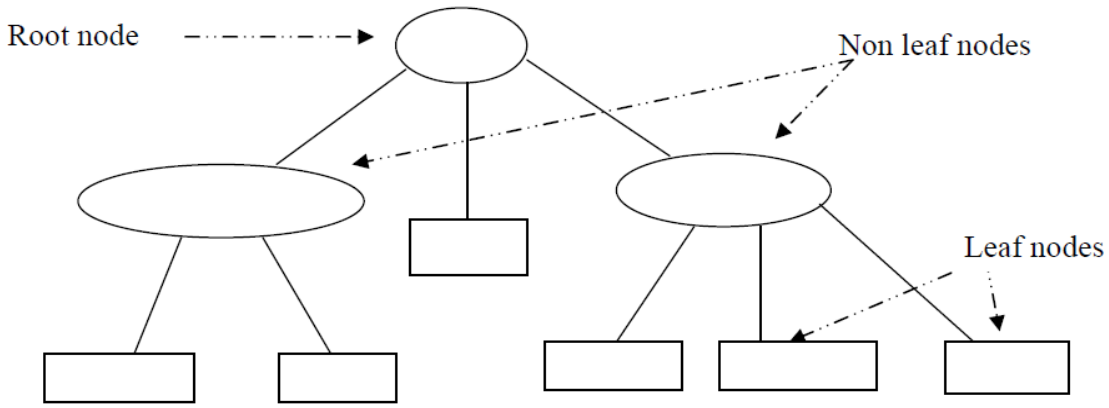

Figure 2.4: Simple decision tree

Decision tree induction is the learning of decision trees or decision tree classifiers from class-labeled training instances. Decision tree classifier is popular and especially attractive due to the following multiple benefits. First, the construction of decision tree classifiers does not require any domain knowledge or providing input parameters, and therefore is appropriate for exploratory knowledge discovery. Second, decision trees can handle high dimensional data. Third, their representation of acquired knowledge in tree form is intuitive and generally easy to comprehend. Fourth, the learning and classification steps of decision tree induction are simple and fast (20). In addition to these, one of the most attractive aspects of decision trees lies in their

interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf. Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule (28).

When decision tree induction is used for attribute subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples. The algorithms that choose the best attribute to partition the data into individual classes include ID3 (Iterative Dichotomize 3), C4.5 (Classification 4.5), and CART (Classification And Regression Tree) (31).

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

There should be certain requirements before decision tree algorithms are applied (31). First: since decision tree algorithms represent supervised learning, they require pre-defined target variables and training dataset which provides the algorithm with the values of the target variable. Second: this training dataset should be rich and varied, providing the algorithm with a healthy cross-section of the types of records for which classification may be needed in the future.

Decision trees learn by example, and if examples are systematically lacking for a definable subset of records, classification and prediction for this subset will be problematic or impossible. Third: the target attribute classes must be discrete i.e. one cannot apply decision tree analysis to a continuous target variable. The target variable needs to take on values that are clearly demarcated as either belonging or not belonging to a particular class.

The challenge with decision tree is overfitting. A problem that can occur is that the model created can overfit the data. Overfitting means that the specification of a model is in large part artifact features of the data set used to build it (i.e., the training set). Overfitting occurs when a

model essentially memorizes the data on which is built. The model should learn the patterns in order to recognize those in future unseen datasets, but the model should not memorize the patterns. The problem with the model memorizing the training set, is that when the model scores an unknown record, it will use the results from the model set if there is a match, and if not, it will produce a random guess. In that case the model is entirely unstable, i.e. it will do no better than random for the score set (24). As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex. This potentially leads to the concept of overfitting which consequently brings the notion of pruning; this implies removing of branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree (30).

Tree Pruning: Decision trees that accurately model the classification in training set will be poor in classifying new cases, i.e., the decision tree is said to be over fit to the training set. With the goal of improving classification accuracy on unseen data, tree pruning attempts to identify and remove branches created from noise and outlier values (18, 30).

Tree pruning can be done in two ways: pre-pruning (or forward pruning) and post-pruning (or backward pruning). Pre-pruning halts the generation of non-significant branches i.e. deciding not to further split or partition the subset of training instances at a given node. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset of instances or the probability distribution of those instances. Post-pruning, on the other hand, first generates the fully grown decision tree and then removes its non-significant branches (27, 30). One such algorithm which uses post-pruning is C4.5 and its successor C5. The C4.5/C5 algorithm first grows an over fit tree and then prunes it back to create a more stable model (16).

The removal of non-significant branches in post-pruning may be accomplished either by sub tree replacement or by sub tree raising. The idea in sub tree replacement is to select some sub tree and replace them with single leaf, basically reducing the number of tests along a certain path. This process starts from the leaves of the fully grown tree and works backward towards the root. In the case of sub tree raising a node may be moved upward towards the root of the tree, replacing the other nodes along the way (18).

Furthermore, a set of classification rules can be extracted from the decision tree by tracing the path from the root to each leaf (corresponding class). This set of rules can be consequently plugged into propitiate knowledge based system (32). So the researcher computed the C4.5 algorithm using J48 method in order to get the best fitted model that can appropriate to predict nutritional status of under-five children in Ethiopia, particularly for the 2011 EDHS dataset and also the investigator tried to generated rules from the J48 decision trees with comparing by the parameter of accuracy measures.

2.3.1.1.2 Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the *naïve Bayesian classifier* to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases (20).

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called *class conditional independence*. It is made to simplify the computations involved and, in this sense, is considered "naïve." *Bayesian belief networks* are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can also be used for classification (30).

Naive Bayes Classifier

Naive Bayesian classifier uses the Bayes' rule to compute the probability of each possible value of the target attribute given the instance, assuming the input attributes are conditionally independent given the target attribute i.e. class conditional independence. Due to the fact that this method is based on the simplistic, and rather unrealistic assumption that the causes are conditionally independent given the effect, this method is well known as Naive Bayes (18, 30). But despite the disparaging name, Naive Bayes works very well particularly when combined with some attribute selection procedure to eliminate redundant (dependent attributes) (18).

The naive Bayesian classifier works as follows:

1. Let *D* be a training set of instances and their associated class labels. As usual, each instance is represented by an n-dimensional attribute vector, $X = (X_1, X_2, \ldots, X_n)$, depicting n measurements made on the instance from n attributes, respectively, $A_1, A_2, \ldots, A_n$.

2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given an instance, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier predicts that instance X belongs to the class C_i if and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \le j \le m; j \ne i$$

Thus probability is obtained for $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem $P(H|X) = \left[\frac{P(X|H)P(H)}{P(X)}\right]$,

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

3. As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots = P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by

$P(Ci) = |C_{i,D}|/|D|$, where $|C_{i,D}|$ is the number of training instances of class C_i in *D*.

4. Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the instance (i.e., there is no dependence relationships among the attributes). Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$$

$$= P(x_1|C_i) * P(x_2|C_i) * \ldots * P(X_n|C_i)$$

We can easily estimate the probabilities $P(x_1)$, $P(x_2|C_i)$, ... , $P(X_n|C_i)$ from the training instances. Recall that here x_k refers to the value of attribute A_k for instance X. For each attribute, we look at whether the attribute is categorical or continuous-valued.

5. In order to predict the class label of *X*, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of instance X is the class C_i if and only if

$$P(X|C_i)\,P(C_i) > P(X|C_j)\,P(C_j) \qquad \text{for } 1 \le j \le m,\ j \ne i.$$

In other words, the predicted class label is the class C_i for which $P(X|C_i)\,P(C_i)$ is the maximum.

The Naive Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which can be used to calculate the probability of each of the possible classifications in turn. Having done this, the class with the largest value will be selected as the class of the new instance (27, 30).

2.3.1.1.3 PART Rule Induction Classification

Decision rule can be constructed from a decision tree simply by following a given path from the root node to any leaf. The complete set of decision rules generated from a class labeled dataset serve the same purpose as decision tree (16). Thus, decision rules are also called as classification rules (27), indicating that the rules can be used to predict the class of an unseen instance.

Rule induction algorithms generate a model as a set of rules logically ANDed together to form the rule antecedent ("IF" part) and the rule consequent ("THEN" part). The antecedent consists of the attribute values from the branches taken by particular path through the tree, while the consequent consists of the class value for the target attribute given by the particular leaf node (16). According to Witten and Frank there are two industrial-strength rule induction algorithms. But the one that works by repeatedly building partial decision trees and extracting rules from them (i.e. PART) is preferred to and used in this research because of its simplicity and its ability to achieve the same level of performance with others (18).

PART algorithm combines the divide-and-conquer strategy (the top-down approach) for decision tree construction with the separate-and-conquer approach for rule learning. The separate-and-conquer strategy first builds a rule and then removes those instances that the rule covers. These consecutive activities continue recursively for the remaining instances until none are left which generates sets of rules called 'decision lists' or ordered set of rules. On the other hand, in the partial decision tree, a pruned decision tree is built for part of the training instances, the leaf with the largest coverage is made into a rule, and the tree is discarded. Using partial decision trees in

conjunction with the separate-and-conquer methodology adds flexibility and speed. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees. During the generation of such a tree, construction and pruning operations are integrated in order to find a "stable" sub tree that cannot be simplified further. Once this sub tree has been found, tree building ceases and a single rule is read off (18).

Finally, both decision trees and decision rules follow same approach to deal with attributes having numeric values. First, numeric values are sorted in descending order and a binary less-than/greater-than test is considered and evaluated in exactly the same way that a binary attribute would be. A threshold i.e. a value that divides the sorted attributes into two equal parts is used to split a non-leaf node (18). In this case the researcher tried to show in comparable performance by implementing several experiments with J48 decision tree algorithm, Naïve Bayes and PART rule induction classifier on 2011 EDHS under-five children dataset.

2.3.2 Descriptive Modeling

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined (33). Descriptive models belong to the realm of unsupervised learning (28). **Unsupervised learning** is a method of learning where a model is fit to observations. It is distinguished from supervised learning by the fact that there is no a priori output. A data set of input objects is gathered and the learner treats them as a set of random variables. A joint density model is then built for the data set. Typically one has a set of N observations $X_1, \ldots, X_N$ having a joint density $p(X)$, all random p-vectors. The goal is to directly infer the properties of this density without the supervising variable. This provides some added difficulty to the characterization. However, it is sometimes an advantage to know that X represents all the variables under consideration and we don't need to infer how $p(X)$ will change, conditioning on the changing values of other variables (29).

Models interrogate the database to identify patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, among others, belong to this family of descriptive models (34).

2.3.2.1 Clustering

Clustering is also referred to as segmentation and it lumps together similar people, things, or events into groups called clusters. It is the identification of classes or clusters for a set of unclassified objects based on their attributes. It is a knowledge discovery process to find groups of interrelated cases and the statistical behaviors that make them adhere into groups. It requires using a distance measure, like the nearest neighbor technique. Once the clusters are decided, the objects are labeled with their corresponding clusters, and common features of the objects in a cluster are summarized to form the class description. For example, a set of new diseases can be grouped into several categories based on the similarities in their symptoms, and the common symptoms of the diseases in a category can be used to describe that group of diseases. Clustering requires significant involvement from a business or domain expert who must judge whether the resulting clusters are useful or not (33).

2.3.2.2 Association rule discovery

Association rule discovery is the process of looking in a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be. In other words, the program takes the initiative in finding what the interesting patterns are, without the user thinking of the relevant questions first. In large databases, there are so many patterns that the user can never practically think of the right questions to ask. The key issue here is the richness of the patterns that can be expressed and discovered and the quality of the information delivered. This in turn determines the power and usefulness of the discovery technique. Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. The rules are given in the form: if item A is part of an event, then X% of the time item B is also part of the event. The rules are written as A_ B, where A is called the antecedent or left-hand side, and B is called the consequent or right hand side. More formally, association rules are of the form A_B, that is, A1, - - - , Am _ B1, - - - - - Bn) Where Ai (for i =1, - - - m} and Bj (for j = 1, - - - - n} are attribute-value pairs (35).

2.4 Methods of Data Mining Tool Selection

The Waikato Environment for Knowledge Analysis (WEKA 3.6.8) manual (36) and Witten and Frank (18) state some of the strengths of the software which can be used to answer these

questions. Therefore, after matching the identified strengths of Weka with the set of criteria as shown in Table 2.2, has passed most of them.

Table 2.2: Evaluation of WEKA 3.6.8 Data Mining Tool

S.No	Criteria/questions	Value
1	What are the ranges of techniques provided by the data mining software?	Preprocessing, Classification, Clustering, Association, attribute selection and Visualize
2	How scalable is the product in terms of the size of the data, the number of fields in the data, and its use of the hardware?	Scalable
3	Does the product provide transparent access to databases and files?	Yes
4	Does the product provide multiple levels of user interfaces?	Yes
5	Does the product generate comprehensible explanations of the models it generates?	Yes
6	Does the product support graphics, visualization, and reporting tools?	Yes
7	Does the product interact well with other software in the environment, such as reporting packages, databases, and so on?	Yes
8	Can the product handle diverse data types?	Yes
9	Is the product well documented and easy to use?	Yes
10	Is there availability of support, training, and consulting?	Yes, but not sure about training
11	How well will the product fit into the existing computing environment?	Compatible
12	Is the vendor/supplier credible?	Yes

As shown in Table 2.2, WEKA 3.6.8 has passed most of the criteria. Therefore, it is selected for creating models and classification rules.

2.4.1 Waikato Environment for Knowledge Analysis (WEKA)

Weka was developed at the University of Waikato in New Zealand. The software is freely available at: http://www.cs.waikato.ac.nz/ml/weka (30). It is open source software released with general public license. This means that Weka is not only free to download and use but its source code is also available to be freely modified and used. The only restriction to all general public license software is that no one has the right to commercially redistribute them which works for Weka also.

Weka is a comprehensive set of advanced data mining and analysis tool. It provides a quick and easy way to explore and analyze data. Weka version 3.6.8, which is latest and stable version, was used in this research. Weka includes two interfaces: command line interface (CLI) and graphical user interface (GUI). Weka contains tools for data pre-processing, classification, regression, clustering, association rules, attribute selection and visualization.

The Weka GUI Chooser

The Weka GUI Chooser provides a starting point for launching Weka's main GUI applications and supporting tools. It includes access to the four Weka's main applications: Explorer, Experimenter, KnowledgeFlow and SimpleCLI.

Figure 2.5: Weka GUI Chooser

A single click of the explorer button opens Weka's explorer window. The window provides graphical representations for activities like preprocessing, attribute selection, learning (association, classification, clustering), visualization. But before the activities listed above are performed, the dataset needs to be imported to Weka from a file which was previously saved in one of the Weka's understandable file formats. Then the filter property has "choose" button which lists several algorithms for preprocessing data. The term filter here is used to refer to algorithms utilized for extracting information about a particular quantity from a set of unclean data (15).

Figure 2.6 shows Weka's explorer window. The window under choose button is partitioned for showing the current relation (its total number of instances and attributes); descriptive statistics

for a selected attribute; listing of all attributes found on the dataset and activities that can be performed on them; and a part of window dedicated to visualize the histogram of selected attribute or a button to visualize all attributes with a new window.

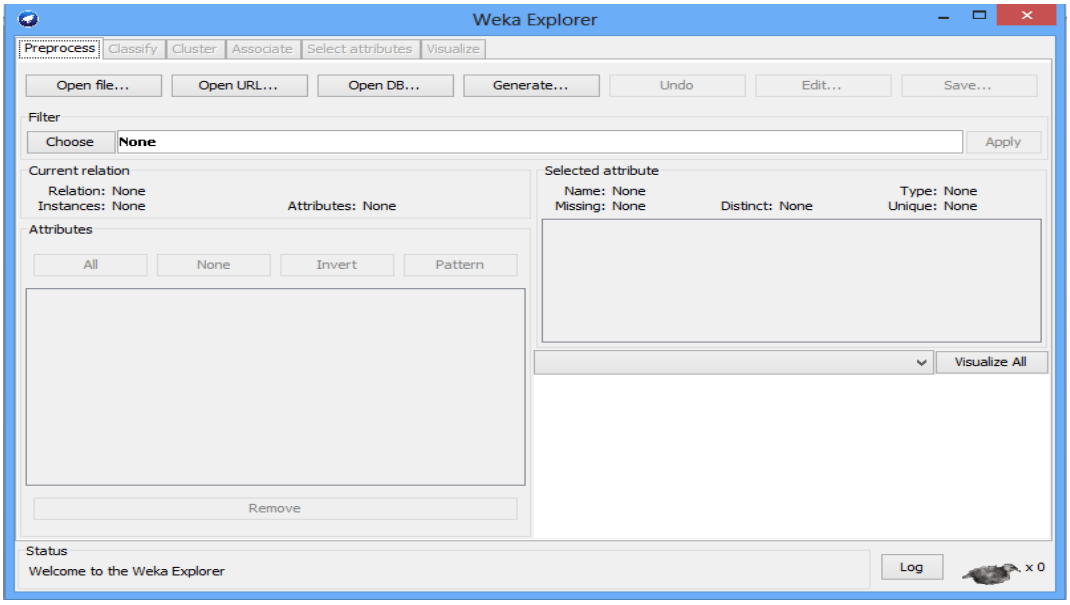

Figure 2.6: Weka's explorer window

Weka Understandable File Formats

Previous versions of Weka require the data presented in a spreadsheet or database to be converted into a Weka understandable format i.e. Attribute Relationship File Format (ARFF). The current version of Weka understands many other file formats including ARFF and Comma Separated Value (CSV). Thus, in this research spreadsheet files were first converted to CSV file format (less effort demanding task) and then opened by using Weka to perform tasks found in data understanding, data preparation and run the data mining algorithms.

2.5 Classifier Accuracy (performance evaluation) Measures

In order to minimize the bias associated with the random sampling of the training and test data samples k-Fold Cross Validation was adopted. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds," D1, D2, : : : , Dk, of approximately equal size. Training and testing is performed k times (34).

As Witten and Frank (18) stated, extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up. Although these arguments are by no means conclusive, and debate continues to range in machine learning and data mining circles about what is the best scheme for evaluation, 10-fold cross-validation has become the standard method in practical terms. Tests have also shown that the use of stratification improves results slightly. Thus the standard evaluation technique in situations where only limited data is available is stratified 10-fold cross-validation.

2.5.1 10-Fold Cross Validation

In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. The classification model is trained and tested 10 times. Each time it is trained on nine folds and tested on the remaining single fold.

10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation. In fact, in data mining community, for methods-comparison studies with relatively smaller datasets, k-fold type of experimentation methods are recommended. In essence, the main advantage of 10-fold (or any number of folds) cross validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as holdout sample (37).

The cross validation estimate of the overall accuracy of a model is calculated by simply averaging the 10 individual accuracy measures

$$CVS = \frac{1}{10}\sum_{1}^{10} Ai$$

Where CVA stands for cross validation accuracy and A is the accuracy measure (e.g. sensitivity, specificity, etc.) of each folds. There are three steps to perform 10-Fold Cross Validation:

Step 1: The complete dataset is randomly divided into 10 disjoint subsets (i.e., folds) with each containing approximately the same number of records. Sampling is stratified by the class labels

to ensure that the proportional representation of the classes is roughly the same as those in the original dataset.

Step 2: For each fold, a classifier is constructed using all records except the ones in the current fold. Then the classifier is tested on the current fold to obtain a cross-validation estimate of its error rate. The result is recorded.

Step 3: After repeating the step 2 for all 10 folds, the ten cross validation estimates are averaged to provide the aggregated classification accuracy estimate of each model type.

2.5.2 Confusion matrix

Confusion matrix is useful tool for analyzing how well classifier recognized the classes. It is body of table with m by m (row and column) matrix the row corresponds to correct classification and the column corresponds to the predicted classifications. An entry, $CM_{i,j}$ in the first m rows and m columns indicate the number of tuples of class that were labeled by the classifier as class j (31). For a classifier to have good accuracy, ideally, most of the tuples would be represented along the diagonal of the confusion matrix with the rest of the entries being closed to zero (17). In confusion matrix, there are classifier evaluation metrics like accuracy, error rate, sensitivity and specificity, precision, recall, and F-measure. Table 2.3 shows two class classification result simple confusion matrix which contains both predicted and actual classes.

Table 2.3: Confusion Matrix with Two Classes Classification

	PREDICTED CLASS		
ACTUAL CLASS		Positive	Negative
	Positive	True Positive (TP)	False Negative(FN)
	Negative	False Positive (FP)	True Negative (TN)

Here are some of performance evaluation computational techniques on confusion matrix that are used in this study. Accuracy is the first one which is widely used to check the performance of the model. It is the percentage of test set tuples that are correctly classified (38).

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

The performance of the model enables it to classify the positive cases correctly is sensitivity. It is defined as the probability of having a positive test result among those with a positive diagnosis for the disease or true Positive recognition rate (38).

True Positive Rate (sensitivity) = TP / (TP+ FN)

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

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

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

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

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

Finally, the F-measure is the inverse relationship between precision & recall (F_1 or F-score**)**: harmonic mean of precision and recall. It is the point to conclude that the precision and recall of the model are significantly balanced (38).

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

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

2.5.3 Receiver Operating Characteristic Curve

A large number of intelligent medical systems (including medical expert systems, neural networks, classifiers, knowledge discovery and data mining systems) showed great progress and they are being developed, practically to aid clinician and to improve patient care in areas such as diagnosis, prognosis, decision support and screening. To test which classifier is highly significant for a given subject is determined by ROC analysis and it becoming widely used tool in medical tests evaluation (39).

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified (40). For example, it can be used to classify adults those who alive and died correctly based on their previous history. The following Figure 2.7 shows the performance of classifier B; that it has the maximum area under curve (12).

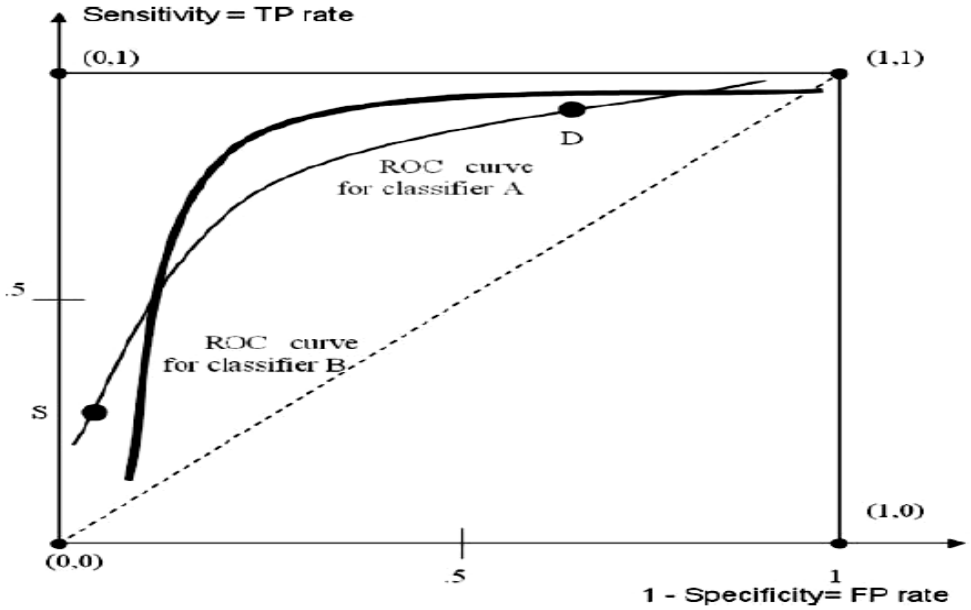

Figure 2.7: Examples for ROC curve

ROC curve is useful visual tool for comparing classification models. It shows the tradeoff between the true positive rate (proportion of positive tuples that are correctly identified) and the false-positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model (30, 12). It is performed by drawing curve in two dimensional spaces by representing vertical axis for true-positive rate and the horizontal axis for false-positive rate (30). In ROC curve, plotting starts at the bottom left-hand corner where the true positive rate and false-positive rate are zero. To plot an ROC curve for a given classification model, one need to rank the test tuples in decreasing order.

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

(30). Thus, from the ROC curve, the closer the ROC curve of a model is to the diagonal line, the less accurate the model is closer to the area of 0.5.

Table 2.4: Performance Measures of ROC Area

ROC Area	**Performance**
0.9-1.0	Excellent(A)
0.8-0.9	Good (B)
0.7-0.8	Fair (C)
0.6-0.7	Poor (D)
0.5-0.6	Fail(F)

The model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger evidence for a positive actual state (1.00) (30, 12, 40). By using ROC analysis one can identify predictors in order to find the one with optimal characteristics and their associated cut-points. Therefore, sensitivity, specificity, precision, F-measure, and ROC area were taken in to account when the classifier performance is evaluated.

2.6 Application of Data Mining in Healthcare

Medicine and health care are among the wide variety of fields where decision is frequently made. In addition to the frequency and urgency needed in decision making, the quality of decision in these areas affect the quality of life directly. Therefore, accurate diagnosis of disease and providing efficient treatment in a timely manner is directly associated with the decision's quality. Towards these human life saving issues, experts assert that appropriate computer-based information and/or decision support systems can aid in achieving accurate clinical test results. In order to pass on quality decision, data mining has a vast potential that can be applied on large volume of data stored in databases (41).

Healthcare also generates large amount of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce mountains of clinical data. This data is a strategic resource for health care institutions (42).

In health care management data mining is applied for variety of issues associated with management. Young and Pita have indicated that United Health Care has mined its treatment instance data with the objective of exploring ways to cut costs and deliver better medicine. The same organization is mentioned as having developed clinical profiles of data mining results to give physicians' information about their practice patterns and to compare it with practices of other physicians and peer-reviewed industry standards (45). Another assurance of data mining applicability in health care management is the clinical best practices initiative of Florida Hospital launched in 1999. The goals of applying data mining in this initiative were developing a standard path of care across all campuses, clinicians, and patient admissions (45).

In relation to drugs, data mining is also found useful to identify adverse drug effects so quickly before they are identified by ordinary methods. Retrospective study conducted by the United States regulatory agency, the Food and Drug Administration, found that a Bayesian statistical analysis (i.e. one algorithm in data mining) of their adverse drug event reporting database would have identified 20 out of 30 known classes of adverse drug events 1-5 years before their detection by existing methods (46). A research has shown that data mining could be utilized to compare the efficiency of different drug regimens for treatment of a particular disease and their cost effectiveness (47).

Researchers have presented an intelligent and effective heart attack prediction methods using data mining. Firstly, they have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weight age, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Five mining goals are defined based on business intelligence and data exploration. The goals are to be evaluated against the trained models. All these models could answer complex queries in predicting heart attack (48)

A study conducted by Tesfahun (44) used for adult mortality prediction at Butajira Rural Health Program open cohort database so as to identify and improve adult health status. The hybrid model that was developed for academic research was followed. Decision tree and Naïve Bayes algorithms were employed to build the predictive model by using a sample dataset of 62,869

records of both alive and died adults through three experiments and six scenarios. Tesfahun overall research design was to build a model that predicts the status of the adult i.e. the probability of he/she alive or die. WEKA 3.6 data mining tools and techniques are utilized as means to address the research problem. In this study as compared to Bayes, the performance of J48 pruned decision tree reveals that 97.2% of accurate results are possible for developing classification rules that can be used for prediction. Further comprehensive and extensive experimentation was needed to substantially describe the loss experiences of adult mortality in Ethiopia.

A study by Soni et.al (48) has applied artificial neural network, naïve bayes, decision tree, and KNN where the decision tree outperformed. The research is conducted with the objective of developing a model of higher accuracy of prediction for heart disease. This research has compared four different supervised machine learning algorithms such as: Naive Bayes, KNN, Decision Tree algorithm, Neural network and reported that Decision Tree outperformed and some time Bayesian classification is having similar accuracy as of decision tree.

Xu Dezhi et.al (49) conducted a research on rule based classification to detect children malnutrition in Sri Lanka. The enhanced the marsma system which was developed under the e-government initiative to provide advice regarding nutrition as well as to provide an easy way for citizens to check nutrition status of their child in cooperating rule based classification technique to detect malnutrition. Further from his research was highlighted that there was an effect on number of rules which is used to make the final decision with the optimality of the final decision.

To summarize, three algorithms namely J48 decision tree, Naïve Bayes and PART rule induction classifiers were used for model building and 10-fold cross validation, Predictive accuracy, TP Rate, TN Rate, Precision, and F-Measure, are six measures used for the evaluation of classification and prediction methods while Predictive accuracy, Weighted TPR, Weighted TNR and Weighted ROC area were used to compare the models.

CHAPTER THREE

MALNUTRITION

3.1 Overview of Malnutrition

Good nutrition is indispensable component of healthy life and access to healthy diet and optimum nutrition are important to good health. Better nutrition means stronger immune systems, less illness and better health. Developing countries such as Ethiopia is experiencing micronutrient malnutrition and macro nutrient under-nutrition. The negative externalities of under-nutrition are many, especially among the younger age group. Nutritional deprivation and infectious diseases among preschoolers feature prominently among the major public health concerns in developing countries. Poor child health and nutrition impose significant and long-term economic and human development costs, especially on the poorest countries and communities, further entrenching their status. Improving child health and nutrition is not only a moral imperative, but also a rational long-term investment (50, 51). Under five years old children are most vulnerable section of the society and this study focuses on these age groups.

3.2 Malnutrition defined

Malnutrition is defined as bad nutrition and can refer to under-nutrition or over-nutrition[1]. Malnutrition manifests itself in many different forms, including: acute malnutrition or wasting; chronic malnutrition or stunting; overweight or obesity; or micronutrient deficiency. Acute malnutrition is the deadliest form of malnutrition and occurs when an individual suffers from current, severe nutritional restrictions, a recent bout of illness, inappropriate childcare practices or a combination of these factors. It is characterized by extreme weight loss and patients must receive supplemental or therapeutic foods in order to recover. Chronic malnutrition, or stunting, can occur when a child suffers from long-term nutrient deficiencies and/or chronic illness, so that not only weight but height is affected. It can also be an outcome of repeated episodes of acute infections, or acute malnutrition. Because it negatively and often irreversibly affects organ growth**,** stunting is strongly linked to cognitive impairment (51).

[1] Overnutrition is part of malnutrition but it is not the scope of this study

Each form of malnutrition increases the risk of disease and early death. Under nutrition, for example, plays a major role in half of all deaths of children under five-years old in developing countries.

3.3 The Causes of Malnutrition

Most causes of child mortality in developing countries are preventable. Malnutrition alone is responsible for over half the under-five-year-old deaths in developing countries, making it one of the most important public health problems in the developing world. Child malnutrition in Ethiopia constitutes a particularly daunting challenge as the country had a 17% under-five mortality rate in 2001, of which an estimated 57% was linked to severe and mild to moderate malnutrition. (11).

As depicted in Figure 3.1, the main different forms of malnutrition may arise due to immediate causes, such as a parasite infection and/or inadequate intake of nutrient, but equally or perhaps even more important are the underlying and basic causes of malnutrition. While disease and an inadequate diet are the immediate causes of malnutrition, underlying causes are of equal importance to examine. Inadequate food security, inadequate care, inadequate health services and unhealthy household environment lead to malnutrition's immediate causes. A lack of potential resources, including financial, human, physical, social or natural, which are triggered by poor social, economic, ideological, and political contexts, result in underlying causes that contribute to malnutrition. The manifestation of the malnutrition depicted in Figure 3.1.

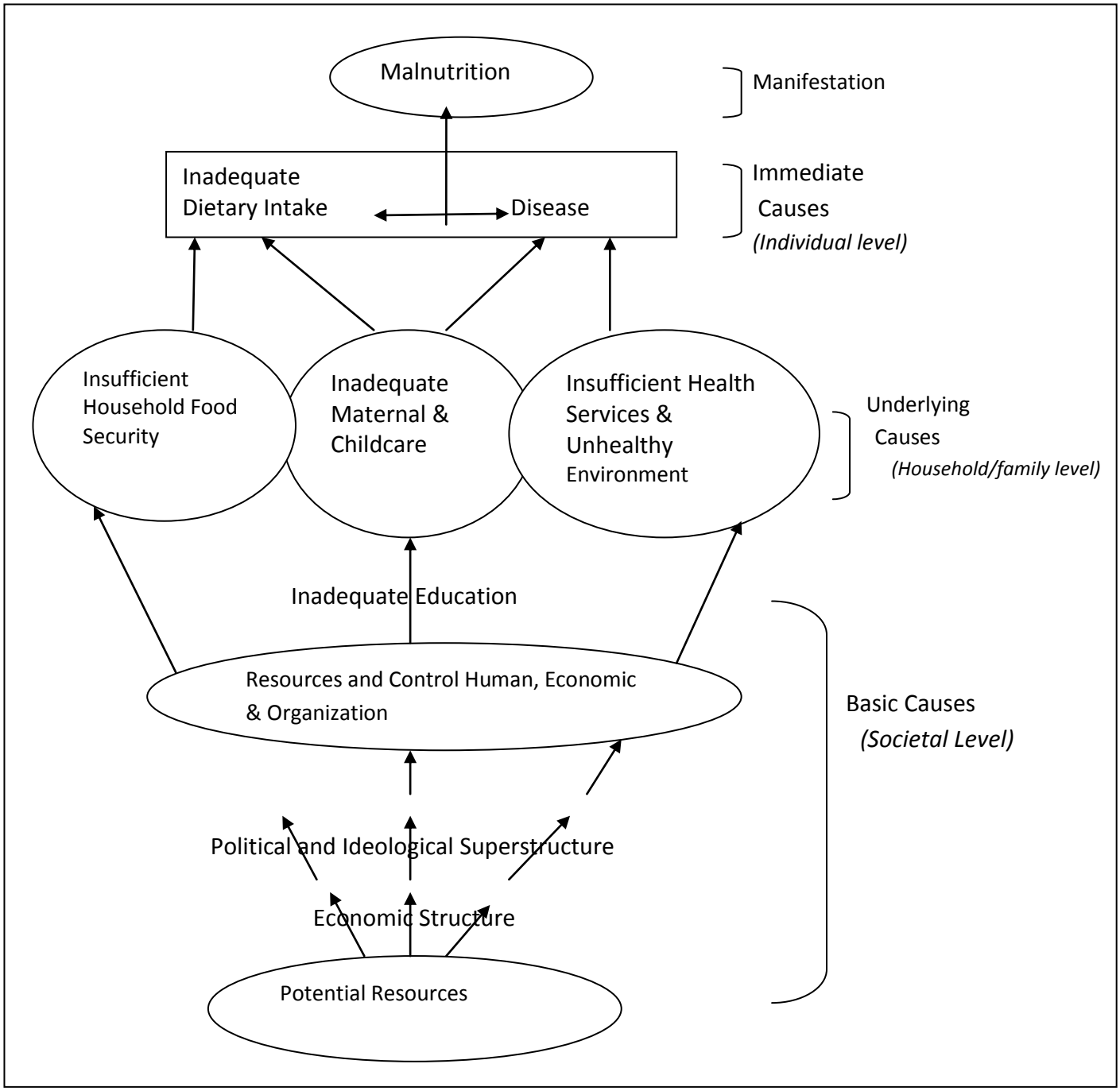

Figure 3.1: Causes of malnutrition (51)

3.4 Nutritional Status of Children

A child's nutritional status reflects the combined effects of many factors, including nutrient intake, health, birth order, and behavioral factors governed by parental preferences. In recognition of the interrelated variables are expressed child's nutritional production function,

they represented as Child's Nutritional status = f (nutritional input, child's health, child's death, births, biological factors, childcare time, technology factors) (50). The nutritional status of children under age five is an important outcome measure of children's health. This evaluation allows identification of subgroups of the child population that are at increased risk of faltered growth, disease, impaired mental development, and death (7).

3.5 Measurement of Nutritional Status of Under-five Children

Indicators of the nutritional status of children are calculated using new growth standards published by the World Health Organization (WHO) in 2006. These new growth standards are generated using data collected in the WHO Multicentre Growth Reference Study. Therefore, the WHO Child Growth Standards can be used to assess children all over the world, regardless of ethnicity, social and economic influences, and feeding practices. The new child growth standards replace the previously used reference standards of the U.S. National Center for Health Statistics, accepted by the U.S. Centers for Disease Control and Prevention (NCHS/CDC/WHO) (50, 51, 7).

Anthropometric indices are used as the main criteria for assessing the adequacy of growth and hence optimal nutritional status in infancy and childhood. Assessment of the nutritional status of the child by nutritional anthropometric indicators of growth has been used not only to provide information on the nutritional and health status of children, but also as an indirect measure of the quality of life of the entire community. Anthropometric method is a quantitative method; it also considers the different types of measurements like, height-for-age z-score (HAZ), weight-for-age z-score (WAZ) and weight for-height z-score (WHZ) (50, 11,7).

Children who fall below minus two standard deviations (−2 SD) from the median of the reference population are regarded as moderately malnourished, while those who fall below minus three standard deviations (−3 SD) from the median of the reference population are considered severely malnourished.

The height-for-age index provides an indicator of linear growth retardation and cumulative growth deficits in children. Children whose HAZ is below −2 SD from the median of the WHO reference population are considered short for their age (stunted), or chronically malnourished.

Children who are below −3 SD are considered severely stunted. Stunting reflects failure to receive adequate nutrition over a long period of time and is affected by recurrent and chronic illness. Height-for-age, therefore, represents the long-term effects of malnutrition in a population and is not sensitive to recent, short term changes in dietary intake (50, 51, 7).

The weight-for-height index measures body mass in relation to body height or length; it describes current nutritional status. Children with WHZ below −2 SD are considered thin (wasted) or acutely malnourished. Wasting represents the failure to receive adequate nutrition in the period immediately preceding the survey and may be the result of inadequate food intake or a recent episode of illness causing loss of weight and the onset of malnutrition. Children with a WHZ below −3 SD are considered severely wasted (50, 51, 7).

The weight-for-height index also provides data on overweight and obesity. Children more than two standard deviations (+2 SD) above the median weight-for-height are considered overweight, or obese (50, 51, 7).

Weight-for-age is a composite index of height-for-age and weight-for-height. It takes into account both chronic and acute malnutrition. A child can be underweight for his/her age because he or she is stunted, wasted, or both. Weight-for-age is an overall indicator of a population's nutritional health. Children with WAZ below −2 SD are classified as underweight. Children with WAZ below −3 SD are considered severely underweight. The Z-score (SD score) is calculated as follows:

$$\text{Z-SCORE} = \frac{\text{measured value} - \text{median of reference population}}{\text{Standard deviation of the refence population}}$$

3.6 Related Works

Under-five children malnutrition remains public health problem in developing countries including Ethiopia. Various studies have been conducted on nutritional status of children to identify the underling causes and associations of malnutrition in Ethiopia.

Teklebrhan (6) has conducted a cross-sectional descriptive study to determine prevalence and predictors of nutritional status of aged 6-36 months children from six randomly selected teaching

health facilities (Jimma, Agaro, Asendabo, Yebu, Serbo health centers) and Limu district hospital within 5-50 km radius from Jimma University. Data were collected through structured questionnaire from 322 mothers/care givers sampled in Jimma zone. The analysis was done by using SPSS for windows version 16.0. Frequency tables were made to see the distribution of variables, cross-tabulations and statistical tests like chi-square test and multivariate logistic regression analysis was made to look for association between nutritional status of children and variables of interest. His study showed that 14.4% were underweight, 33.9% were stunted and 19.2% were wasted in the area.

Beka et.al (10) have conducted a community based cross-sectional survey on 622 mother-child pairs of 0-59 months old children to identify magnitude and determinants of stunting in under five years of age in food surplus region of Ethiopia in the case of Mecha and Wenberma Woredas of West Gojam Zone, Northern Ethiopia between May and June 2006. Data were collected using a pre-tested structured questionnaire interviews. Both bivariate analysis and multivariate analysis (logistic regression model) were used to identify the determinants of under-five stunting.

The statistical analyses were carried out using SPSS 12.01 for windows and Anthropometric indices were calculated using Epi-Info 6.0 software. According to Beka et al. the analyses revealed that 43.2% of the children under age five were suffering from chronic malnutrition, 14.8% were acutely malnourished and 49.2% were found to be under-weight.

Aweke et.al (4) has conducted a cross-sectional survey with descriptive and analytical components on nutritional status of children in food insecure households in two districts i.e, Lalomama and Gerakeya of North Showa zone of Amhara in 2007. Two hundred food insecure households were selected from two districts. Anthropometric and clinical data were also collected from a total of 239 (151 < 5yrs and 88, 6-12yrs) children. Data on demographic, childcare, feeding practices and morbidity status of children were collected using an interview, community focus group discussion and secondary data from district offices. The Epi-Info version 3.4.3 and SPSS 12.0 for windows were used to enter and analyze data. The overall prevalence of stunting, underweight and wasting was 54.2%, 40.2% and 10.6 %, respectively in the districts.

Abdu (52) conducted a study to explore the effect of maternal characteristics on the health and nutritional status of under-five children using the 2005 EDHS dataset. Abdu used a survey of 14,070 women with ages ranging from 15 to 49 and 6033 men with ages from 15 to 59 across all eleven geographic (administrative) areas (nine regions and two city administrations) of the country. During the survey, 5280 under-five children were identified in the households. The health and nutritional status of children are measured using the two widely used anthropometric indicators HAZ and WHZ. The statistical analysis was carried out using SPSS 16. In the ordinary least squares estimation, it is observed that maternal characteristics have a significant impact on child health and nutritional status. The magnitudes of the coefficients, however, were found to slightly increase when maternal education is instrumented in the two-stage least-squares (2SLS) estimation. Moreover, in the quantile regression estimation, the impacts of maternal characteristics were observed to vary between long-term and current child health and nutritional status.

Shegaw (53) applied data mining techniques to investigate the potential applicability of data mining technology to predict the risk of child mortality based up on community based epidemiological dataset gathered by the Butajira Rural Health Program epidemiological study. Shegaw used a sample dataset consisting 1,100 records taken randomly from the two classes of children (i.e. alive and died) of the ten years surveillance dataset study which contains a total of 64,077 records. To build predictive models he used neural network and decision tree techniques and the performances were 93% and 95% respectively. He stated that decision tree approach provided simple rules that can be used by nontechnical health care professionals to identify cases for which the rule is applicable.

Be'emnetu (54) also applied data mining techniques to investigate the potential applicability of data mining technology to predict under-five mortality based up on community based epidemiological dataset gathered by the Butajira Rural Health Program epidemiological study. Be'emnet used hybrid six step processes on ten years surveillance dataset of the Butajira Rural Health Program epidemiological study which contains a total of 11,600 records. To build predictive models he used Decision tree and Naïve Bayes techniques and the performances were 97.49% and 96.67% respectively. He stated that decision tree approach provided simple rules

that can be used by nontechnical health care professionals to identify cases for which the rule is applicable.

Biset (55) conducted a research on application of data mining to predict low birth weight using EDHS 2005 data set with the purpose of identifying the determinant factors affecting low birth weight. The methodology employed to perform the research work is CRISP-DM. Weka software was used to extract the hidden patterns among the variables under the study. The selected data mining techniques for predicting low birth weight was classification on 9,861 records. To build predictive models she used J48 decision tree classifier and PART rule induction algorithms were selected for experiments and the performances were 94.35% and 94.7% respectively. In general, the results from this study were encouraging; it can be used as decision support aid for health practitioner. The extracted rules in both the algorithms are very effective for the prediction of low birth weight

The above reviews revealed that a wide variety of issues in the health sector are making use of the potentials of data mining. This study applies data mining for the improvement of decision making in the area of predicting nutritional status of under-five children in Ethiopia. It is also experiments algorithms such as J48, Naïve Bayes and PART rule induction. However, the capability of the algorithms to work in multiclass situation and understandability of models they provide are among the additional reasons for choosing these algorithms.

CHAPTER FOUR

UNDERSTANDING AND PREPROCESSING 2011 EDHS DATASET

It is well known that the success of every data mining research is strongly dependent on the quality of data processing, business understanding and data understanding. Data pre-processing could be critical and a very complicated task. Sometimes, the data pre-processing takes more than half of the total time spent on solving the data mining problem because incomplete, noisy, and inconsistent data are common place properties of large real-world databases and data warehouses (56, 57). Thus, data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to clean the noisy data, extract and merge the data from different sources, and then transform and convert the data into a proper format (58). This chapter describes data understanding and data preprocessing in the following sections.

4.1 The Raw Data Description

The source data employed for this research purpose is 2011 EDHS dataset. This dataset is collected from 2006/2007-2010/2011. The 2011 EDHS was conducted under the support of the MoH and CSA. The census is conducted in every five years intervals. The primary objective of the 2011 EDHS was to provide up-to-date information for policy makers, planners, researchers and programme managers, which would allow guidance in the planning, implementing, monitoring and evaluating of population and health programmes in the country. The various sector development policies and programmes assist and monitor the progress towards meeting the Millennium Development Goals (7).

The 2011 EDHS collected information on the population and health situation which covers family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, malaria, and women's empowerment.

The 2011 EDHS dataset has eight data (i.e, children's records, birth records, couple's records, HIV test record, household member records, individual records and male records) in SPSS file format. Each data include other attributes such children's data. In children's data, there are household attributes, male attributes, birth attributes and couples attributes with the total of 920

attributes. The 2011 EDHS is a nationally representative survey of 11, 654 instances on children collected in order to classify nutritional status, anemia level, and others in the study from women age of 15-49, men age of 15-59 years and under five year children on 920 attributes. This sample provides estimates of health and demographic indicators at the national and regional levels, and for rural and urban areas. Among all under-five survey data, child data attributes such as sex of child, age of child, height of child, weight of child, Height for Age Z-score (HAZ), Weight for Age Z-score (WAZ), Weight for Height Z-score (WHZ), anemia level (hemoglobin level), etc are included. Mother's background characteristics are mother's age, region, place of residence, literacy, BMI, religion, wealth index, mother's education, occupation, hemoglobin level (anemia), etc. On the other hand, the data also included household (HH) characteristics such as, age of -HH, sex of HH, education of HH, occupation of HH, etc.

From original children's data other's attributes and unrelated attributes with nutritional status of under-five children removed. Finally the selection of the dataset is performed by the help of literature and domain experts. List of source dataset, 2011 EDHS dataset, before preprocessing is depicted in Annex1.

4.2. Data Understanding

Data understanding phase mainly focuses on creating a target dataset with selected sets of variables that is relevant to the discovery process. Without understanding the existing data, it is difficult to draw the target dataset from the original since the world data is unclean and not appropriate at the source to run mining process (12).

The original dataset from SPSS is exported to excel file because Weka data mining tool does not accept SPSS format and whose size amounted to 14.6 MB before any processing activity was done on it. Under-five data which found in electronic format has 11,654 instances and 920 attributes. This 920 attributes are not only on nutritional status but on HIV, vaccination, breast feeding, child preference, nutrition (under-five, adult and women), etc.

The aim of this study is to create a model based on secondary data that was selected on the base of nutritional status of under-five children in Ethiopia. The entire attributes in the original dataset were not concerned for this experimentation. Thus, only relevant attributes were considered so as

to achieve the objective of the study. From total of 920 attributes which were found in under-five children records, 44 attributes which related with nutritional status were selected. From total 44 attributes, 14 repeated attributes, 8 least important attributes and 6 attributes those have more than 50% missing values were minimized. Table 4.1 shows that repeated attributes.

Table 4.1 Repeated attributes

Field name	Data Type	Descriptions
V437	Numeric	Respondent's weight in kilo grams (1 decimal)
V438	Numeric	Respondent's height in centimeter (1 decimal)
V439	Numeric	Height/Age percentile
V439	Numeric	Height/Age standard deviation
V444	Numeric	Weight/Height percent ref.median (WHO)
V444A	Numeric	Weight/Height standard deviation (DHS)
M19	Numeric	Birth weight in kilograms (3 decimals)
M19A	Numeric	Weight at birth/recall
HW2	Numeric	Child's weight in kilograms (1 decimal)
HW3	Numeric	Child's height in centimeters (1 decimal)
HW6	Numeric	Height/Age percent ref.median
HW9	Numeric	Weight/Age percent ref.median
HW12	Numeric	Weight/Height percent ref.median
HW73	Numeric	BMI standard deviation (new WHO)

Table 4.1 shows repeated attributes due to repetition with selected variables. For example, BMI of the child shows that nutritional status. But to check nutritional status of a child anthropometric index has been taken in the study. Height and weight of the child used to calculate HAZ, WAZ and WHZ. The same true for listed attributes in Table 4.1.

Table 4.2: Least important attributes

Field name	Data Type	Descriptions
V131	Nominal	Ethnicity
V150	Nominal	Relationship to house hold
V151	Nominal	Sex of house hold head
V155	Nominal	Literacy of the mother
V453	Scale	Mother's hemoglobin level
V457	Nominal	Mother's anemia level
B5	Nominal	Child is alive
HW53	Scale	Child hemoglobin level (g/dl - 1 decimal)

These attributes depicted in Table 4.2 were minimized based on the literature and domain expert support. For instance, literacy of the mother shows that educational status. This also included in

the selected attributes. Due to the above mentioned reason eight attributes minimized from the total selected attributes.

Table 4.3: More than 50% missing values

Field name	Data Type	Descriptions	Total Missing
V130	Nominal	Religion	51.6%
V463A	Nominal	Smokes cigarette	60.1%
V729	Nominal	Husband/partner's educational attainment	50.5%
V730	Numeric	Husband/partner's age	54%
V733	Nominal	Husband/partner's occupation	70%
M34	Scale	When child put to breast	53.9

During test and explorative analysis, six attributes found more than 50% missing values. In data mining research attributes those have more than 50% missing values can affected the whole result of the prediction. From this angle, researcher minimized these six attributes from the total attributes.

Finally, source data was not labeled or clustered by nutritional status (new class). Based on the 2006 world health organization children multicenter standards, researcher labeled nutritional status.

4.3 Data Preparation

Data preparation is the most important phases of the data analysis activity which involves the construction of the final data set (data that will be fed into the modeling tool) from the initial raw data. Data preparation generates a dataset smaller than the original one, which can significantly improve efficiency of data mining. This task includes: attribute selection, filling the missed values, correcting errors, or removing outliers (unusual or exceptional values), resolve data conflicts using domain knowledge or expert decision. On an effort to make the dataset used for this study suitable few data transformation methods were used. Discretization was used to reduce distinct values of attributes, dimensionality reduction was used to reduce the size of the dataset and attribute selection method was applied to remove weakly relevant attributes.

4.3.1 Attribute Selection

Deciding on the data that was used for the analysis was based on several criteria, including its relevance to the data mining goals as well as quality and technical constraints such as limits on data volume or data types (59). Therefore, in this study the attributes are selected with the help of domain expert and extensive literature review because taking all the variables in the database we have, feed them to the data mining tool and find those which are the best predictors may be does not work very well. One reason is that the time required to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models (28). Thus, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts and literature review in order to simplify the task of modeling.

The national survey data set obtained contains many attributes. To decide on the relevant attributes for this study, the researcher has discussed with domain experts in the area. As described in Table 4.1, the following attributes are selected from the five years survey: Mother's age, Mother's educational level, Mother's BMI, Mother's occupation, Residence, Region, Wealth quintile, Size of child at birth, child's age, child's sex, child's HAZ, child's WAZ, child's WHZ, anemia level, total number of children and ever had vaccinated. The final selected attributes were prepared and preprocessed as stated in the following section, before developing the models.

The description of selected attributes, data types they take, the unit of measure used, list of values or range of values of these attribute are given together with statistical summaries of these attributes in data description and exploratory data analysis section were discussed as follows. The main objective of analyzing statistical summaries of these selected attributes is to see the distribution of each value of attributes in the dataset to identify errors (noises) and detect whether there exist missing values or not.

Table 4.4: Description of the selected attributes from 2011 EDHS Dataset

Field name	Type	Descriptions
MOTHERAGE	Numeric	Mother's age
REGION	Nominal	Region
RESIDENCE	Nominal	Type of place of Residence
MOTEDUC	Nominal	Mother's Education level
WEALTHINDEX	Nominal	Wealth index
TOTAL	Numeric	Total children ever born
MOTHBMI	Numeric	Mother's Body mass index
MOTHEROCCUP	Nominal	Respondent's occupation (grouped)
SEXOFCHILD	Nominal	Sex of child
AGEOFCHILD	Numeric	Current age of child
SIZEOFCHILD	Numeric	Size of child at birth
EVERHADVAC	Nominal	Ever had vaccination
CHILDANEM	Nominal	Child anemia level
HAZ	Nominal	Height/Age standard deviation (new WHO)
WAZ	Nominal	Weight/Age standard deviation (new WHO)
WHZ	Nominal	Weight/Height standard deviation (new WHO)
NUTSTATUS	Nominal	Nutritional status of under-five (WHOCGMS)

All attributes have relation with nutritional status of under-five children and explained briefly as follows.

MOTHER AGE: Mother's age is one of the attributes in relation with nutritional status of child. For data management purpose, mother's age categorized into seven groups (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49) with five intervals based on literatures.

REGION: Region grouped into nine regions and two city administrations (i.e. Tigray, Afar, Amhara, Oromiya, Somali, Benishangul-Gumz, Southern Nation Nationality People (SNNP), Gambela, Harari, Addis Ababa and Dire Dawa) of mother and child respectively.

RESIDENCE: Living place with the value of Rural or Urban.

MOTEUC: Indicator of mother's education level with four classifications; No education, Primary, Secondary, Higher values.

WEALTHINDEX: Wealth index/quartile indicator of wealth level of household with value of Poorest, Poorer, Middle, Richer, Richest.

TOTAL: It indicates total number of children with value of 1-2 children, 3-4 children, 5-6 children, >6 children in the family.

MOTHBMI: Mother Body Mass Index (BMI) is indicator of nutritional status of mother's. Attribute values are <18.5 (Thin mother), 18.5-24.9 (Normal mother) or >25 (Over/Obese mother).

MOTHOCCUP: indicator of mother's occupation status whether Working of Not working.

SEXOFCHILD: Sex of child with value of Male or Female.

CHILDAGE: Age of child in months with six categorical (<6, 6-11, 12-23, 24-35, 36-47, 48-59) values.

SIZEOFCHILD: Size of child at birth in kilogram with the value of <2.5kg (small), 2.5-4kg (Normal), >4kg (large).

EVERHADVAC: Ever had vaccinated attributed shows a child vaccinated or not vaccinated.

CHILDANEM: Child anemia level also indicated nutritional level in addition to anthropometric indicators and health level with the value of severe, moderate, mild, not anemic.

HAZ: Height-for-age Z-score shows stunted or long term malnutrition level of a child with value of less than minus two standard deviation (<-2SD), normal (-2SD to 2SD) and over/obse (>2SD).

WAZ: Wight-for-age Z-score shows underweight malnutrition level of a child with value of less than minus two standard deviation (<-2SD), normal (-2SD to 2SD), and over (>2SD).

WHZ: Weight-for-height Z-score shows wasting (acute or current malnutrition level) of a child with value of less than minus two standard deviation (<-2SD), normal (-2SD to 2SD), and over (>2SD).

NUTSTATUS: This new created attribute, nutritional status of under-five children is a class in data mining language and categorized or labeled based on the WHO Growth Multicenter standards of 2006 with class of Normal, Wasted, Underweight, Stunted.

4.3.2 Selection of Instances

In addition to the removal of irrelevant attributes which were done based on the attributes; relevance to the prediction of nutritional status of under-five children, instances that deal with nutritional status were selected from the dataset. Out of the 11,654 under-five data, 9,607 remain. As this study uses classification algorithms for the purpose of predictive model building, 846 records without class information are removed from subsequent analyses because of died children. This is because building a predictive model requires to give the learner algorithm with a training set that have all instance whose outcome or dependent attribute (class label) is not missing. Instance with missing values for outcome class are not useful for predictive model building in data mining because classification algorithms of data mining learn how instances are

classified under the different classes. The classes do not exist, means the algorithm learns nothing from these instance. Records without class labels (missing or not entered) should be ignored, provided that the data mining task involves classification (30). Even, from this number, 1201 records are also removed due to over-nutrition, i.e out of objective of this study (i.e, under nutrition). The remaining dataset has 9607 records whose outcomes are distributed in one of the outcome categories (normal, wasted, underweight, stunted).

4.3.3Exploratory Data Analysis

In this section, efforts have been made to present the description of the selected attribute together with the exploratory data analysis performed with the use of frequency tables. The attribute's description, data type, unit of measure and list of values or range of values are described. With the use of frequency tables, the exploratory data analysis was performed to detect bad data i.e. attributes with the missing values and wrong entries or noises and inconsistency in values of attributes. The frequency tables for the selected attributes show the original distribution of values of attributes in instances of the dataset before any preprocessing is done on the dataset.

Mother's Age: Mother's age is important demographic variable in the study of nutritional status. The age of mothers is classified by five year age groups. This attribute is categorized into seven groups: 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 as shown in Table 4.5.

Table 4.5: Summary of Mother's Age Attribute

MOTHER'S AGE			
N	Category	Frequency	Percent (%)
Valid	15-19	393	4.1
	20-24	1893	19.7
	25-29	3019	31.4
	30-34	1996	20.8
	35-39	1487	15.5
	40-44	615	6.4
	45-49	204	2.1
Missing		0	0.00
Total		9607	100

As depicted in Table 4.5, most of the mothers (31.4%) are at age group of 25-29 and small numbers of mothers (2.1%) are from 45-49. No missing values and another work need at mother's age attribute.

Region: The region attribute indicates the location of mothers and child. This attribute contains a total of 11 administrative region of the country. The distinct values of region attribute are Tigray, Afar, Amhara, Oromiya, Somali, Benishangul-Gumz, Southern Nation Nationality People (SNNP), Gambela, Harari, Addis Ababa, Dire Dawa. Table 4.6 shows the distribution of mothers and child by region.

Table 4.6: Summary of Region Attribute

REGION			
N	Category	Frequency	Percent (%)
Valid	Tigray	1077	11.2
	Affar	891	9.3
	Amhara	1068	11.1
	Oromiya	1515	15.8
	Somali	755	7.9
	Benishangul-Gumuz	834	8.7
	SNNP	1367	14.2
	Gambela	682	7.1
	Harari	511	5.3
	Addis Ababa	323	3.4
	Dire Dawa	584	6.1
Missing		0	0.00
Total		9607	100

From Table 4.6, it can be seen that most of the data (15.8%) are collected from Oromiya region and small data (3.4%) collected from Addis Ababa. No missing values and other inconsistencies are found in the data.

Place of Residence: Place of residence is nominal attributes; the possible values of this attribute are urban and rural. The value of this attribute is described in Table 4.7.

Table 4.7: Statistical Summary of residence Attribute

RESIDENCE			
N	Category	Frequency	Percent (%)
Valid	Urban	1544	16.1
	Rural	8063	83.9
Missing		0	0.00
Total		9607	100

As we can see from Table 4.7, the majority (83.9%) of respondents reside in rural areas and y small (16.1%) group live in urban area.

Mother's Education: This attribute reveals the level of education of a mother. Mother's education is indirectly related to a child's health. Mother's education is nominal attribute that contains four distinct values (No education, Primary, Secondary, and Higher) as shown in Table 4.8.

Table 4.8: Statistical summary of levels of Mother's Education Attribute

MOTHER'S EDUCATION			
N	Category	Frequency	Percent (%)
Valid	No education	6695	69.7
	Primary	2446	25.5
	Secondary	307	3.2
	Higher	159	1.7
Missing		0	0.00
Total		9607	100

According to Table 4.8, the most frequent value for educational level of the mothers is no education (69.7%) and few mothers (1.7%) attended higher education.

Wealth Index: It serves as an indicator of level of wealth that is consistent with expenditure and income measures. The index was constructed using household asset data via a principal components analysis. The distinct values of wealth index attributes are poorest, poorer, middle, richer and richest as shown in table 4.9.

Table 4.9: Statistical summary of mother's wealth index attribute

WEALTH INDEX			
N	Category	Frequency	Percent (%)
Valid	Poorest	2959	30.8
	Poorer	1776	18.5
	Middle	1592	16.6
	Richer	1583	16.5
	Richest	1697	17.7
Missing		0	0.00
Total		9607	100

From Table 4.9, it can be seen that most mothers (30.8%) of mothers are the poorest and some (16.5%) are richer. No missing values exist in total instances of wealth index.

Total number of children: This is an indicator for number of children born in the family. The distinct values of total number of ever born attributes are 1-2, 3-4, 5-6, and >6 children as shown in Table 4.10.

Table 4.10: Statistical summary of total number of ever born children attribute

TOTAL NUNBER OF CHILD EVER BORN			
	Category	Frequency	Percent (%)
Valid	1-2	2855	29.7
	3-4	2804	29.2
	5-6	1978	20.6
	>6	1970	20.5
Missing		0	0.0
Total		9607	100

According to Table 4.10, most of the families have 1-2 (29.7%) and 3-4 (29.2%) children. Also some families have 5-6 (20.6%) and more than 6 (20.5%) children.

Mother's BMI: Mother's MBI is an indicator for nutritional status of a mother. The distinct value of mother BMI attribute is under or thin (<18.5kg), normal (18.5-4kg) and/or over (>4kg) as shown in Table 4.11.

Table 4.11: Statistical summary of mother's BMI attribute

MOTHER'S BMI			
N	Category	Frequency	Percent (%)
Valid	<18.5	2520	26.2
	18.5-24.9	6458	67.2
	>=25	604	6.3
Missing		25	0.3
Total		9607	100

Table 4.11 depicts that the distribution of instances among the values of the mother's BMI, majority are normal (67.2%) and few of them are over (6.3%). The frequency of missing value with code 9999 in the dataset amounts 50 (0.3%) of the total instances.

Mother's occupation status: Mother's occupation is an indicator of economic level in relation to nutritional status of a child. The distinct values of occupation attribute is not working and working as shown in Table 4.12.

Table 4.12: Statistical summary of mother's occupation

MOTHER'S OCCUPATION			
N	Category	Frequency	Percent (%)
Valid	Not working	5168	53.8
	Working	4439	46.2
Missing		0	0.0
Total		9607	100

Like most of the attributes selected for nutritional status predictive model building, the number (frequency) of instances are unevenly distributed for the values of mother's occupational status, i.e. 53% and 46.2% are not working and working respectively.

Size of a child at birth: Size of a child at birth reveals nutritional status of a child in the house hold. The distinct values of this attributes are large (>4kg), normal (2.5-4kg) and small (<2.5kg).

Table 4.13: Statistical summary of the size of a child at birth

SIZE OF CHILD AT BIRTH			
N	Category	Frequency	Percent (%)
Valid	Small	2771	28.8
	Normal	3792	39.5
	Large	3018	31.4
Missing		26	0.3
Total		9607	100

Size of a child at birth (in kilogram) is almost normally distributed in the instances. 0.3% are missing from total instances, which are very small and may distort the information that the attribute provides for the algorithms during model building.

Ever had vaccination: is an attribute used to check that a child is vaccinated or not. The distinct values of this attribute is not vaccinated and vaccinated as shown in Table 4.14.

Table 4.14: Statistical summary of ever had vaccination

CHILD EVER HAD VACCINATION			
N	Category	Frequency	Percent (%)
Valid	No	1975	20.6
	Yes	5236	54.5
Missing		2396	24.9
Total		9607	100

Table 4.14 depicts that most of the children had vaccination (54.5%) and some of them (20.6%) are not vaccinated. The attribute has the largest (24.9%) missing values among all attributes in the dataset. These missing values need additional work manually or using application software.

Child anemia level: This attribute is an indicator for anemia level of a child and an indicator for nutritional status of a child. The distinct values of this attributes are severe, moderate, mild and not anemic as shown in Table 4.15.

Table 4.15: Statistical summary of child anemia level

CHILD ANEMIA LEVEL			
N	Category	Frequency	Percent (%)
Valid	Severe	298	3.2
	Moderate	2051	21.3
	Mild	1785	18.6
	Not anemic	4164	43.3
Missing		1309	13.6
Total		9607	100

Table 4.15 shows the majority (43.3%) of the children are not anemic but some of them (3.2%) are severely anemic. There are some (13.6%) missing values that need additional work.

Sex of a child: Sex of a child is nominal attribute with possible values of male and female. The statistical distribution of this attribute is shown in Table 4.16.

Table 4.16: Statistical summary of sex of a child

SEX OF CHILD			
N	Category	Frequency	Percent (%)
Valid	Male	4888	50.9
	Female	4719	49.1
Missing		0	0.0
Total		9607	100

Table 4.16 depicts that the distribution of instances among the values of the sex of a child is almost even; male (50.9%) and female (49.1%). There is no missing value for the attribute.

Child Age: This is the age of a child from 0-59 months. The possible values of this attribute are: <6, 6-11, 12-23, 24-35, 36-47 and 48-59 months as shown in Table 4.17.

Table 4.17: Statistical summary of children's age category

CHILD AGE			
N	Category	Frequency	Percent (%)
Valid	<6	1001	10.4
	6-11	967	10.1
	12-23	1775	18.5
	24-35	1876	19.5
	36-47	2054	21.4
	48-59	1934	20.1
Missing		0	0.00
Total		9607	100.0

Table 4.17 depicts that the distribution of the instances among the values of the children's age group of 36-47 (21.4%) and 6-11 (10.1%) are the large and small age category, respectively.

Nutritional Status: nutritional status of children under age five is an important outcome measure of children's health. The anthropometric data on height and weight permit the measurement and evaluation of the nutritional status of the children. Initially there was no variable labeled with nutritional status (class) but it contains anthropometric measurements, i.e, Height for Age Z-score (HAZ), Weight for Age Z-score (WAZ) and Weight for Height Z-score (WHZ). Nutritional status classification has been done by the researcher based on WHO Growth Multicenter standards of 2006. Clustered nutritional status attribute values are stunted, underweight, wasted and normal as shown in Table 4.18.

Table 4.18: Statistical summary of nutritional status attribute

NUTRITIONAL STATUS OF UNDER-FIVE			
N	Category	Frequency	Percent (%)
Valid	Normal	4713	49.1
	Stunted	2339	24.3
	Underweight	1449	15.1
	Wasted	1106	11.5
Missing		0	0.0
Total		9607	100

Table 4.18 depicts that the distribution of instances among the values of the nutritional status categories. This evaluation allows identification of subgroups of the children population that are at increased risk of faltered growth, disease, impaired mental development, and death. From the total instances, 49.1% are normal and 11.5% are wasted. No missing values because of those children without anthropometric indices removed intentional by the researcher.

4.4 Data Preprocessing for Mining

Proceeding to data mining step with low-quality data will lead to low-quality results (6). Thus, the dataset is preprocessed to improve the quality of the data and therefore to get good data mining models. Data preparation or preprocessing is done with the objective of cleaning the data from quality problems such as missing values.

4.4.1 Managing Missing Values

Missing values refers to one or more fields of an attribute which have no value in it. The existence of many such cases makes datasets incomplete and building models of any type, whether descriptive or predictive, with incomplete data makes the resulting model non representative of the reality (12). As it was learnt from data understanding step, with the use of descriptive statistical summaries, some of the attributes are having missing values. Thus, the missing values found under each attribute in the attributes selected for this study are replaced automatically by a feature called "ReplaceMissingValues" in Weka. "ReplaceMissingValues" replaces the mode of nominal valued attribute and the mean of continuous valued attribute for missing values. Replacing the mode or the mean is a preferred method to removing an instance only because of a single missing value in a particular cell (30). Table 4.19 shows the attributes' percentage of missing values that applied by "ReplaceMissingValues" implements.

Table 4.19: The percentage of missing values for the selected attributes.

o	**Attributes**	**Missing values (%)**
1	Ever had vaccination	24.9
2	Mother's BMI	0.3
3	Size of child at birth	0.3
4	Child anemia level	14

As shown in Table 4.19, missing values found in four selected attributes are small, except ever had vaccination attributes (24.9%) however, not insignificant. Therefore, these missing values are managed automatically by replacing them with the most frequent value (mode in statistical language) using weka 3.6.8 tool.

4.4.2 Data Transformation

Once the data has been assembled and major data problems are fixed, the data must still be transformed for analysis. This involves adding derived fields to bring information to the surface. It may also involve smoothing, aggregation, generalization, normalization, discretization, and attribute construction (37).

Discretization is the process of converting continuous valued variables to discrete values where limited numbers of labels are used to represent the original variables. The discrete values can have a limited number of intervals in a continuous spectrum, whereas continuous values can be infinitely many (38). Here discretization is made on three anthropometric indices (HAZ, WAZ and WHZ) based on 2006 WHO Growth Multicenter standards and size of child at birth.

4.4.2.1 Discretizing the Values of HAZ Attribute

HAZ: The attribute HAZ index provides an indicator of linear growth retardation and cumulative growth deficits in children. Final HAZ attribute discretized categories are -2SD to -3SD moderately malnutritioned,and <-3 severely malnutritioned. Finally these two values discretized into one value, stunted. -2SD to 2SD (normal) and >2SD (over) as shown Table 4.20.

Table 4 20: Statistical summary of HAZ attribute

HAZ			
N	Category	Frequency	Percent (%)
Valid	<-2SD	4070	42.37
	-2SD-2SD	5262	54.77
	>2SD	275	2.86
Missing		0	0.00
Total		9607	100

Table 4.20 depicts HAZ of under-five children whose HAZ is below −2SD from the median of the WHO reference population are considered short for their age (stunted), or chronically malnourished. Children who are below −3SD are considered severely stunted. Stunting reflects failure to receive adequate nutrition over a long period of time and is affected by recurrent and chronic illness. HAZ, therefore, represents the long-term effects of malnutrition in a population and is not sensitive to recent, short term changes in dietary intake. Based on this information,

most of literature and domain expert advised to be merged or discretized as stunted. Most of (54.77%) under-five children are normal and few of them are over nutrition (2.86%).

4.4.2.2 Discretizing the Values of WAZ Attribute

WAZ: Attribute WAZ is a composite index of height-for-age and weight-for-height. It takes into account both chronic and acute malnutrition. Final WAZ attribute discretized categories are <-2SD (underweight), -2SD to 2SD (normal) and >2SD (over) as shown in Table 4.21.

Table 4.21: Statistical summary of WAZ attribute

WAZ			
N	Category	Frequency	Percent (%)
Valid	<-2SD	2884	30.0
	-2SD-2SD	6663	69.4
	>2SD	60	0.6
Missing		0	0.0
Total		9607	100

From Table 4.21, a child can be underweight for his/her age because he or she is stunted, wasted, or both. WAZ is an overall indicator of a population's nutritional health. Children with weight-for-age below minus two standard deviations re classified as underweight. Children with WAZ below −3SD are considered severely underweight. Based on this information, most of literature and domain expert advised to be discretized as underweight. Most of (69.4%) under-five children are normal and few of them are over malnutritioned (0.6%).

4.3.2.3 Discretizing the Values of WHZ Attribute

WHZ: The WHZ index measures body mass in relation to body height or length; it describes current nutritional status. Final WHZ attribute discretized categories are <-2SD (Wasted), -2SD to 2SD (Normal) and >2SD (Over) as shown Table 4.22.

Table 4.22: Statistical summary of WHZ attribute

WHZ			
N	Category	Frequency	Percent (%)
Valid	<-2SD	1128	11.7
	-2SD-2SD	8315	86.6
	>2SD	164	1.7
Missing		0	0
Total		9607	100

As shown in Table 4.22, children with WHZ below −2SD are considered thin (wasted) or acutely malnourished. Wasting represents the failure to receive adequate nutrition in the period immediately preceding the survey and may be the result of inadequate food intake or a recent episode of illness causing loss of weight and the onset of malnutrition. Children with a WHZ index below −3 SD are considered severely wasted. Based on this information, most of literature and domain expert advised to be merged or discretized as wasted. Most of (86.6%) under-five children are normal and few of them are over malnutritioned (1.7%).

4.3.2.4 Discretizing the Values of Size of Child at Birth Attribute

For easy interpretation, high level concept is generated for size of child at birth attribute. The values are very large and larger than average is more than 4kg, average (normal) is equal to 2.5-4kg and smaller than average and very small is less than 2.5kg. This is done by defining a portion of the values through explicit data grouping as presented in Table 4.20. It shows that discretization of size of a child at birth. Except second one, Normal, the first two values merged into large and last two values merged into small.

Table 4.23: Size of child at birth attribute values generated by explicit data grouping

Old Values	New Values
Very Large	Large
Larger than Average	
Average	Normal
Smaller than Average	Small
Very Small	

4.5 Description of Preprocessed and Prepared Data

At the beginning of this chapter the dataset acquired as an information source is described. Since then, different activities were performed on the dataset with the objective of making it suitable for the data mining algorithms and producing representative model. Very large numbers of instances were removed and large numbers of attributes are removed. Different corrective measures were applied on the remaining attributes. The final summary of the dataset ready for experiments is shown in Table 4.24.

Table 4.24: Summary of the selected dataset

Parameters	**Original dataset**	**Target dataset**	
Total Number of Records	11,654	9,607	
Total Number of attributes	920	17	
File Format	SPSS 16.0	.xls	.csv
Size of Data	14.6 MB	2.42 MB	1.14MB

The dataset whose general description is given in Table 4.24 is ready to be imported to the data mining tool (Weka 3.6.8). After importing the comma separated values (CSV) and attribute related file format (ARFF) files to Weka with the use of "explorer" interface, experiments explained in the next chapter have been conducted to meet the objectives of the study.

CHAPTER FIVE

EXPERIMENTATION AND EVALUATION OF DISCOVERED KNOWLEDGE

In this study an attempt was made to explore 2011 EDHS dataset to predict nutritional status of under-five children. The purpose of the experiments in classification is to find a model that is able to predict the nutritional status of under-five children in Ethiopia as normal, wasted, underweight, or stunted taking selected variables as inputs. The experiments conducted in this study are predictive model building experiments.

5.1 Experimental Design

Cleaned dataset is used for predictive model building. All the experiments that are discussed in the subsequent sections are carried out using 9607 instances and 17 attributes. The attribute set includes HAZ, WAZ, WHZ, MOTHAGE, REGION, RESIDENCE, MOTHEDUC, WEALTHINDEX, MOTHBMI, MOTHOCCUP, TOTAL, SEX, CHILDAGE, SIZE, CHILDANEM, EVERHADVAC and NUTSTATUS. The last attribute in the list represents the class attribute which is mandatory in developing predictive models i.e. the dependent variable in statistics language.

In order to build predictive models for nutritional status, three different data mining algorithms were applied. The three algorithms are selected based on their good performance during test and they have literature support. More specifically, J48 decision tree, Naïve Bayes and PART rule induction are the algorithms with which predictive model building experiments are conducted. In 10 fold cross validation, one option in Weka for the purpose mentioned; the dataset is split into 10 equal parts. The model is trained on nine–tenth of the dataset and then the classifier is tested on one-tenth. This way, the error of the resultant model will be the average of all the models found during each fold or iteration.

The algorithms used for predictive model building experimentations are found in Weka 3.6.8. This version works on many file formats than its predecessors and it is compatible with CSV file

format. Thus, no additional effort was exerted to change the dataset from excel to ".arff" file format which is necessary in the previous versions. The prepared dataset is saved using CSV file format. Then, this file is imported to Weka.

In order to build a model csv file is given to the classifiers. The frontend of this (NutStatusOfU5InEthiopia.csv) file is shown in Figure 5.1.

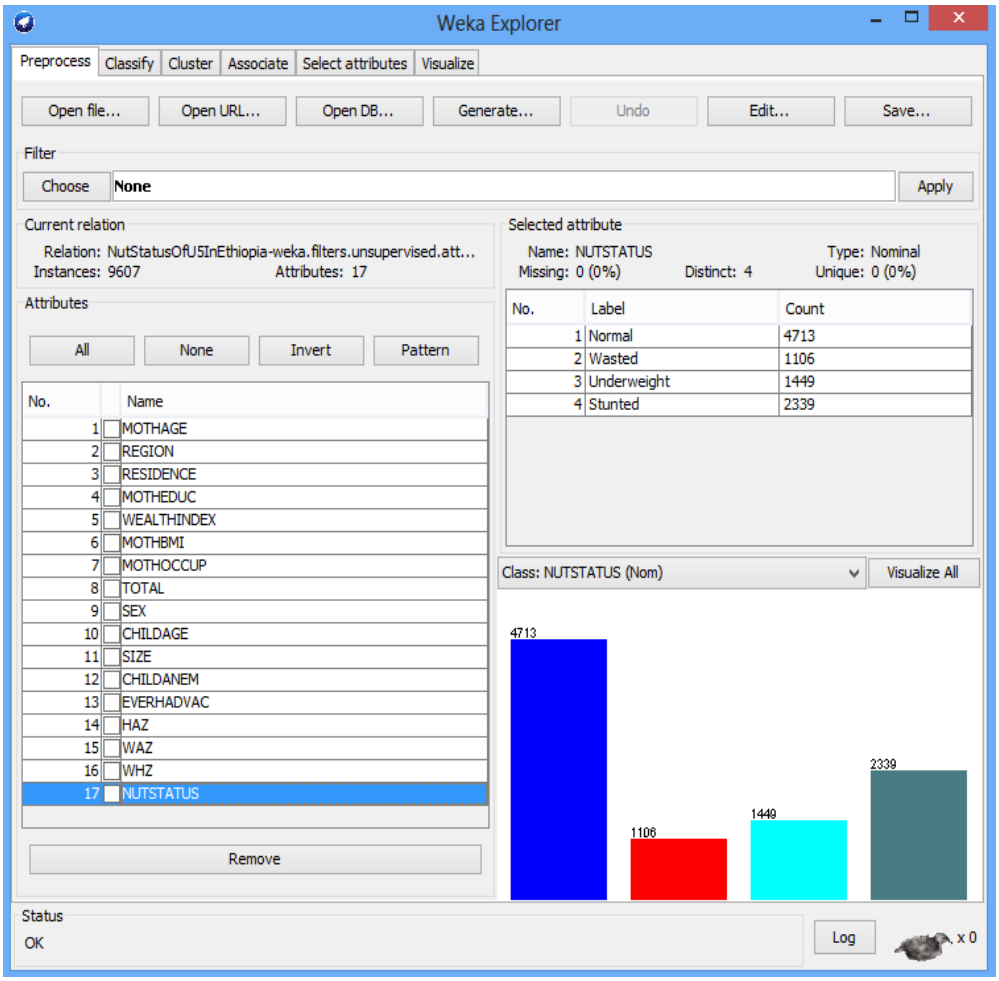

Figure 5.1: Weka 3.6.8 explorer window showing the list of selected attributes

As it has been seen in Figure 5.1, the four classes, 9607 cases consists of 4713(49.1%) normal, 1106 (11.5%) wasted, 1449 (15.1%) underweight, and 2339(24.3%) stunted, dataset were applied for experimentation. Figure 5.1 shows that the dataset is imbalanced before prediction and model development.

A dataset is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. There is a special technique called SMOTE to solve such type of problems in dataset. It is also the case that the cost of misclassifying an abnormal (interesting) example as a normal example is often much higher than the cost of the reverse error. Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. Frequently classifiers are developed using class-imbalanced data, i.e., data sets where the number of samples in each class is not equal. Standard classifications methods used on class-imbalanced data often produce classifiers that do not accurately predict the minority class; the prediction is biased towards the majority class. The methods of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC area) than only under-sampling the majority class (60). Over-sampling the minority class and under-sampling the majority class can achieve better classifier performance (in ROC area) than varying the loss ratios. There are imbalanced classes observed from the Figure 5.1, majority classes and minority classes.

Table 5.1: Imbalanced classes before and after SMOTE

Class	Before SMOTE	100% SMOTE	200% SMOTE	300% SMOTE	400% SMOTE
Normal	4713	4713	4713	4713	9426
Wasted	1106	2212	2212	2212	2212
Underweight	1449	1449	2898	2898	2898
Stunted	2339	2339	2339	4678	4678

As shown in Table 5.1, there were imbalanced classes (one majority and three minority classes) occurred. The main problem before SMOTE was ignorance of the tools during rule generation i.e, adding instances to the majority classes. For more information see Annex 2. The other problems in first (100%) and second (200%) SMOTE case was increasing instances only for

wasted and underweight classes. The forth (400%) SMOTE also deviated rule of SMOTE (impossible to SMOTE majority class).

To avoid the effect of data imbalance on the model created, weka based SMOTE technique is applied. This is automatic operation where minority classes are over sampled by generating synthetic examples of minority class and adding them to the dataset to make the target attribute balanced (60). If the class attribute is imbalance, this condition further needs balancing by different techniques. Unless the classes are proportional, the classification will be skewed in dominant classes. Consequently, the new predicted instances will also fall in the dominant classes erroneously unless the classes' proportionality is considered (16).

Therefore, assuming such kind of problem in this case, SMOTE technique on the dataset selected for both training and testing have been applied three times (300%) using Weka data mining tool. Here, the researcher has taken 300% SMOTE as the threshold because after the third experiment oversampling the minorities will lead to under sampling of previously majority classes, despite the continuous decrease in accuracy and continuous increase in ROC area.

Thus, the classification accuracy of the minority class become increased in the SMOTE technique for certain level i.e. the total of 14501 instances (4713 normal, 2212 wasted, 2898 underweight and 4678 stunted) were provided and the subsequent experimentations were conducted based on this sample dataset. Figure 5.2 shows the original dataset and 300% (three times) balanced one after applying the SMOTE technique.

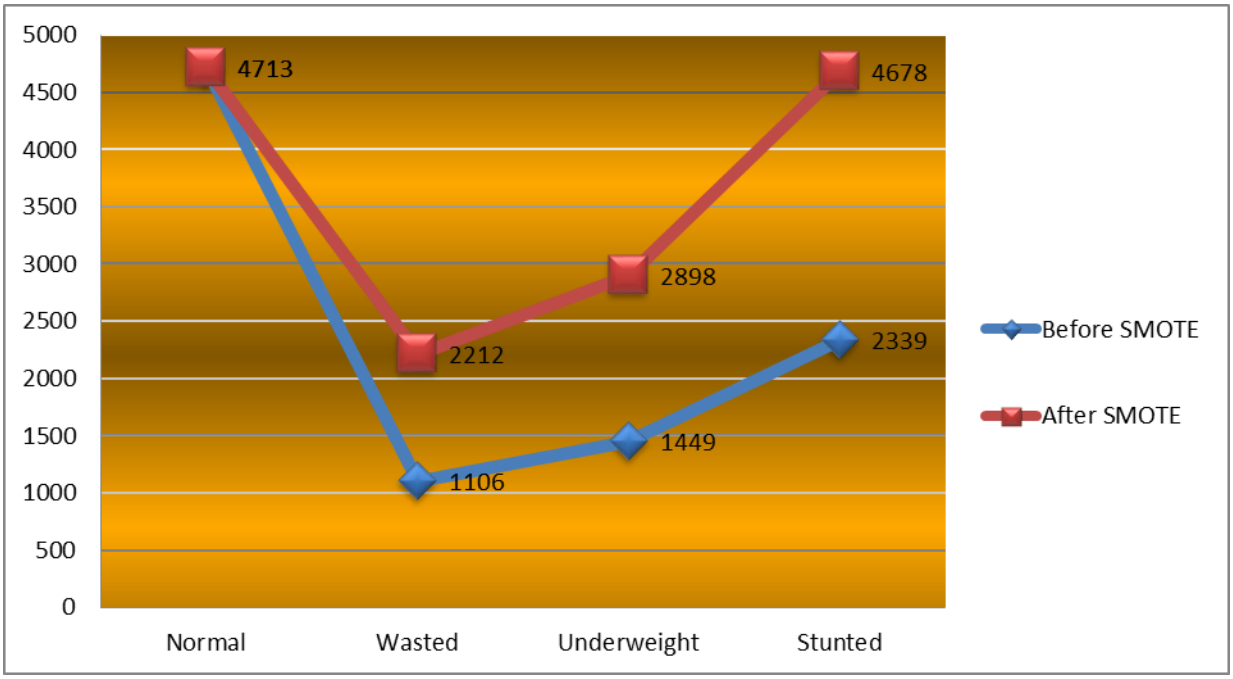

Figure 5.2: Review of the original class variable using SMOTE

5.2 Selecting and Evaluating the Attributes

Attributes selection involves searching through all possible combinations of attributes in the data to find which subset of attributes work best for prediction. To do this, "InfoGainAttributeEval" and "ChiSquareAttributeEval" methods were used to assign a worth to each subset of attribute by searching ranker style in the Weka. Table 5.2 shows rank of selected attributes.

Table 5.2: Attributes evaluation by InformationGainAttributeEval and ChiSquareAttributeEval

Rank #	Name Of Attributes	Informationgain	Rank#	Name Of Attributes	ChiSquareAttributeEval
1	HAZ	0.73816	1	WHZ	14268.19418
2	WHZ	0.60394	2	HAZ	11368.24155
3	WAZ	0.56478	3	WAZ	8751.37242
4	CHILDAGE	0.08292	4	CHILDAGE	1577.62869
5	REGION	0.047	5	REGION	938.16191
6	WEALTHINDEX	0.03077	6	WEALTHINDEX	632.96984
7	MBMI	0.02089	7	MBMI	490.24783
8	MATHEDUC	0.02094	8	MATHEDUC	415.93063
9	RESIDENCE	0.01894	9	RESIDENCE	394.25737
10	SIZE	0.01603	10	SIZE	321.46004
11	CHILDANEM	0.01447	11	CHILDANEM	293.00299
12	EVERHADVAC	0.00805	12	EVERHADVAC	163.18881
13	MOTHAGE	0.00728	13	MOTHAGE	145.82242
14	MOTHOCCUP	0.00708	14	MOTHOCCUP	141.46341
15	TOTAL	0.00308	15	TOTAL	62.28584
16	SEX	0.00139	16	SEX	27.85793

Table 5.2 depicts attribute selection and evaluation using by attribute selection tools methods “InformationGainAttributeEval” and “ChiSquareAttributeEval”. Difference observed between HAZ and WHZ attributes. In selecting using “InformationGainAttributeEval” method, HAZ comes first whereas using “ChiSquareAttributeEval”, WHZ comes first. In weka, attributes by “ChiSquareAttributeEval” >150 (for example attributes 1-12 in Table 5.2) considered as good association with class (in this case, nutritional status).

5.3 Algorithm Classifier Parameters

J48 Classifier Parameter Options

During experimentations, all the default parameters that already set in the Weka were used except that the unpruned ‘False’ was changed to unpruned ‘True’ during J48 unpruned tree generation in order to experiment the model performance without pruning the tree.

Table 5.3: J48 Classifier Parameter Options

Parameters	Descriptions	Parameter type
binarySplit	When to use binary split on nominal attributes when building the trees.	Boolean
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning).	Numeric
Debug	If set to true, classifier may output additional info to the console.	Boolean
minNumObj	The minimum number of instances per leaf.	Numeric
NumFolds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	Numeric
reducedEroorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning.	Boolean
saveInstanceData	Whether to save the training data for visualization.	Boolean
Seed	The seed used for randomizing the data when reduced-error pruning is used.	Numeric
subtreeRaising	Whether to consider the sub tree raising operation when pruning.	Boolean
Unpruned	Whether pruning is performed.	Boolean
usedLaplace	Whether counts at leaves are smoothed based on Laplace.	Boolean

BinarySplits parameter by default is set to “False”. If this value is changed to “True”, it enforces the model generated to be binary decision tree rather than generalized decision tree. The confidence factor helps to set a limit so that the algorithm makes more or less pruning. The default value for confidence factor is 0.25. Smaller values of confidence factor enforce more pruning. The working of confidence factor requires the unpruned parameter to be set to “False”. The subtreeRaising parameter is by default set to “True” to replace the nodes in a decision tree

with a leaf during pruning. Experiment #2, 4, 8 and #10 in Table 5.9 were done by setting unpruned parameter value to "True". The tree generated will represent unpruned decision tree.

Naïve Bayes classifier parameter

The second algorithm applied in this research is NaiveBayes with all variables and selecting variables. The most important parameter in relation to this study is displayModelInOldFormat. However, there are also other parameters which can be adjusted according to needs of data used in different research areas. The default value to this parameter is "False". The researcher has altered this value to "True" as displaying the model in old format is recommended to output the classifier's result for multi-valued class classification after SMOTE data.

PART Rule Induction classifier parameter

The third data mining technique used in this research is PART Rule induction algorithm. PART algorithm extracts rules. Due to this reason the algorithm is categorized under classification by rule induction. The detailed procedure of this algorithm in extracting rules is explained in chapter two. The algorithm builds partial decision trees and reads a path from the root of the tree to the leaf to read of a rule. The rules are added together to give a complete set of rules. PART has almost a similar set of parameters with J48 algorithm that can be adjusted to build better model from datasets. Table 5.4 shows some of PART rule learner parameters.

Table 5.4: Summary of the PART rule induction parameter

Parameters	**Description**	**Types**
binarySplits	Whether to use binary splits on nominal attributes when building the partial trees	Boolean
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning)	Numeric
minNumObj	The minimum number of instances per rule.	Numeric
reducedErrorPruning	Whether reduced-error pruning is used instead of C4.5 pruning	Boolean
unpruned	Whether pruning is performed	Boolean

The parameter "binarySplits" is left as it is by default, since many of the attributes, including the class, in the current study are multi valued. The default value of unpruned False" parameter

algorithm is adjusted by the researcher to experiment on the dataset as "True". Other parameters of the algorithm are left as they are by default.

5.4 Model Building

To build the model, ten different experiments were conducted using J48 decision tree, Naïve Bayes classifier and PART rule induction. All experimentations during model building applied after SMOTE (300%). Table 5.5 shows model selections and experimentation.

Table 5.5: Experiments and Scenarios

Exp#	Experiments	Scenarios (Models)
1	J48 pruned Tree Model Generation	J48 pruned tree with all 16 attributes
2	J48 unpruned Tree Model Generation	J48 unpruned tree with all 16 attributes
3	J48 pruned Tree Model Generation	J48 pruned tree with 12 selected attributes
4	J48 unpruned Tree Model Generation	J48 unpruned tree with 12 attributes
5	Naïve Bayes Classifier	Naïve Bayes with all 16 attributes
6	Naïve Bayes Classifier	Naïve Bayes with 12 selected attributes
7	PART pruned rule induction	PART pruned with all 16 attributes
8	PART unpruned rule induction	PART unpruned with all 16 attributes
9	PART pruned rule induction	PART pruned with 12 selected attributes
10	PART unpruned rule induction	PART unpruned with 12 selected attributes

The intention here is to investigate the effect of all attribute on classification accuracy as well as model ROC area, complexity and decision tree size on both pruned and unpruned J48 tree classifiers. According to "ChiSquareAttributeEval" attribute ranker in weka tool, the first 12 or 1-12 attributes selected from table 5.2. During test J48 and PART rule algorithm classifiers except, Naïve Bayes classifier, showed slight decreases in accuracy and ROC area with selected attributes. Therefore, J48 and PART rule induction classifiers applied using all 16 attributes. The second algorithm, Naïve Bayes classifier was evaluated on model performance with all and selected attributes. The third algorithm, PART rule induction, was evaluated its accuracy, ROC area and rules size on both pruned and unpruned rule induction with all attributes. Before conducting different experiments, sample dataset was learnt by assigning different dataset for each type of set in model creation and model usage.

5.4.1 Experimentation with J48 Algorithm

J48 is Weka's implementation of the C4.5 algorithm which can work on multiple valued attributes. As it was observed from the data description, the attributes that affect nutritional status are multi valued. In addition to using the default parameter settings of the algorithm to build predictive model with J48, an attempt was made to find better classifier by varying its important parameters. Table 5.6 shows that J48 pruned experimentations with all attributes.

Table 5.6: Experimentation with J48 Decision Tree

Exp#	Model	SMOTE	Acc	WTPR	WFPR	WROC
1	J48 pruned tree with all attributes	Before	92.6%	92.6%	1.5%	0.984
2	J48 pruned tree with all attributes	100%	93.3%	93.3%	1.3%	0.985
3	J48 pruned tree with all attributes	200%	94.2%	94.2%	1.8%	0.981
4	J48 pruned tree with all attributes	300%	92.2%	92.2%	2.6%	0.973

Table 5.6 shows dataset before and after SMOTE using J48 decision tree algorithm. Before SMOTE J48 decision tree algorithm cannot assign instances in to classes due to data imbalances. Imbalance data need SMOTE one time (100%), two times (200%) and three times (300%) on the original dataset. During first SMOTE changes made on only wasted class, second SMOTE for only underweight class and third SMOTE for only stunted class. The fourth SMOTE oversampled majority class (Normal). According to literature, the SMOTEd classes must not over majority class (Normal class). Based on this SMOTED data, third (300%) was selected for experimentation.

5.4.2 Experimentation with Naïve Bayes Algorithm

Bayesian methods are based on the assumptions of probability. The Naïve Bayes algorithm assumes the attributes are independent. The probability of co-occurrence of an attribute value together with a particular outcome value is computed. Then, the class of a new instance will be computed by multiplying the probabilities of values the instance has assumed under each attribute. Table 5.7 depicts experimentation with Naïve Bayes with all attributes.

Table 5.7: Experimentation with Naïve Bayes Classifier

Exp#	Model	SMOTE	Acc	WTPR	WFPR	WROC
1	Naïve Bayes with all attributes	Before	92.2%	92.2%	1.6%	0.986
2	Naïve Bayes with all attributes	100%	93.1%	93.1%	1.3%	0.989
3	Naïve Bayes with all attributes	200%	93.1%	93.8%	1.9%	0.984
4	Naïve Bayes with all attributes	300%	89.7%	89.7%	2.8%	0.976

Table 5.7 shows dataset before and after SMOTE using Naïve Bayes algorithm. Before SMOTE Naïve Bayes algorithm cannot classify instances in to classes due to data imbalances. Imbalance data need SMOTE one time (100%), two times (200%) and three times (300%) on the original dataset. During first SMOTE changes made on only wasted class, second SMOTE for only underweight class and third SMOTE for only stunted class. The fourth SMOTE oversampled majority class (Normal). Based on this SMOTED data, third (300%) was selected for experimentation.

5.4.3 Experimentation with PART rule induction Algorithm

PART rule induction algorithm can work on multiple valued attributes. In addition to using the default parameter settings of the algorithm to build predictive model with PART, an attempt was made to find better classifier by varying its important parameters. Table 5.8 shows that PART pruned experimentations with all attributes.

Table 5.8: Experimentation with PART rule induction

Exp#	Model	SMOTE	Acc	WTPR	WFPR	WROC
1	PART pruned rule with all attributes	Before	91.1%	91.1%	2.2%	0.978
2	PART pruned rule with all attributes	100%	92.3%	92.3%	1.3%	0.981
3	PART pruned rule with all attributes	200%	93.2%	93.2%	2%	0.98
4	PART pruned rule with all attributes	300%	92.6%	92.6%	2.5%	0.978

Table 5.8 shows dataset before and after SMOTE using PART rule induction algorithm. Before SMOTE PART rule induction algorithm cannot classify instances in to classes due to data imbalances. Imbalance data need SMOTE one time (100%), two times (200%) and three times (300%) on the original dataset. During first SMOTE changes made on only wasted class, second SMOTE for only underweight class and third SMOTE for only stunted class. The fourth SMOTE oversampled majority class (Normal). Based on this SMOTED data, third (300%) was selected for experimentation.

Based on the experimental design, Table 5.9 depicted that experimentation of three algorithms through changing their schemes.

Table 5.9: Experimentation after 300% SMOTEd with three selected algorithms

Exp#	Model	Accuracy	WTPR	WFPR	WROC
1	J48 pruned tree with all attributes	92.24%	92.2%	2.6%	0.973
2	J48 unpruned tree with all attributes	92.73%	92.7%	2.6%	0.966
3	J48 pruned tree with selected attributes	90.58%	90.6%	2.4%	0.975
4	J48 unpruned tree with selected attributes	91.68%	91.7%	2.9%	0.973
5	Naïve Bayes with all attributes	89.68%	89.7%	2.8%	0.976
6	Naïve Bayes with selected attributes	89.94%	89.8%	2.8%	0.976
7	PART pruned rule with all attributes	**92.62%**	**92.6%**	**2.5%**	**0.978**
8	PART unpruned rule with all attributes	92.82%	92.8%	2.27%	0.96
9	PART pruned rule with all attributes	91.72%	91.7%	2.8%	0.976
10	PART unpruned rule with all attributes	90.84%	90.8%	3.3%	0.971

***Key**: **EXP#**= Experiment Number, **WTPR**= Weighted Average True Positive Rate, **WFPR**= Weighted Average False Positive Rate and **WROC**= Weighted Average Receiver Operating Characteristics area.*

Table 5.9 revealed that exp# 1-4 done by J48 pruned and J48 unpruned with all and selected attributes. Both experiments have good results in terms of accuracy and WROC area. Applying the second algorithm, Naïve Bayes exp# 5&6 has the same WROC area, 97.6% and 97.6%

respectively but little difference in terms of accuracy (i.e 89.68% and 89.94%) achieved. The four experimentations (7-10) using PART pruned and PART unpruned rule induction with all attributes has good performance. In terms of WROC area, PART pruned rule induction have scored the highest (97.8%) in comparison with J48 and Naïve Bayes classifiers. According to literature, SMOTEd dataset evaluated based on ROC area. ROC shows tradeoff between TPR (sensitivity) and FPR (specificity). Due to SMOTEd dataset applied, PART pruned rule induction model is used to extract interesting rules or patterns based on its high performance in its ROC area.

5.4.3 Performance Evaluation and comparison of Classifiers

One of the objectives of this study was to compare and evaluate the techniques which were used in the study, such as decision tree, Naïve Bayes and PART rule induction classifiers and to select the one, which performs the best. To evaluate and compare the performance of each of the classifiers involved in this study, the standard metrics of accuracy, True-Positive Rates, False-Positive Rates and ROC are applied.

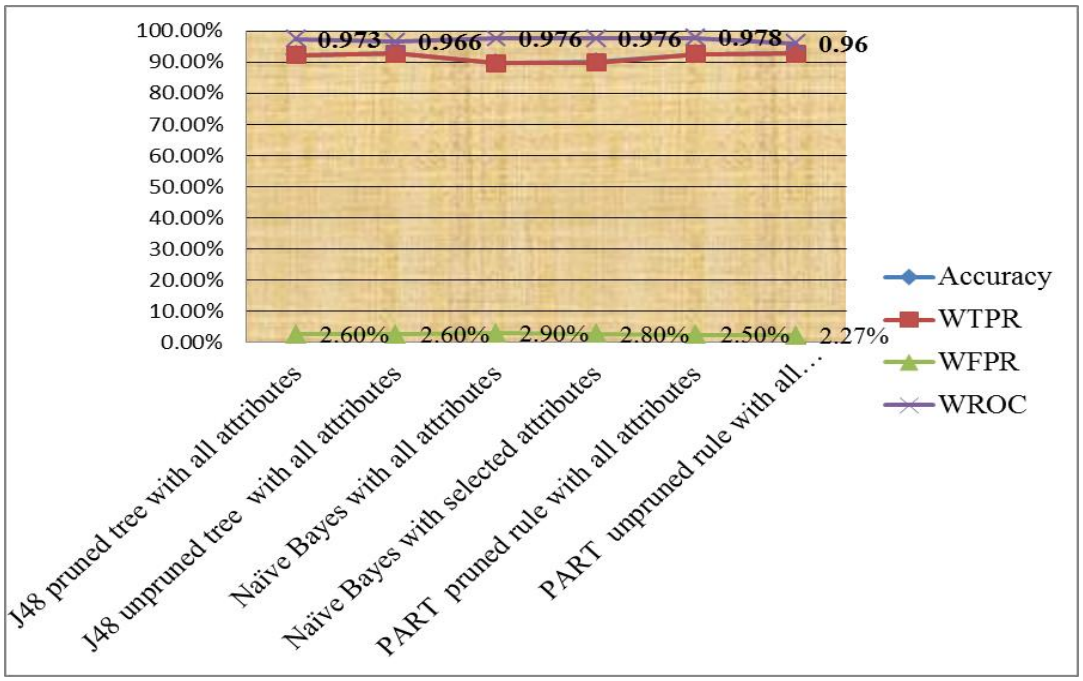

Figure 5.3: Models Performance Comparison

As shown in Figure 5.3, the experiments are mainly categorized in to models that involve the name of the algorithm used and their corresponding parameters. All the experiments were done after three times (300%) SMOTE. In medical and health researches accuracy and ROC area

performance is compared. Various results have been obtained using J48 pruned tree, J48 unpruned tree, Naïve Bayes models, and PART rule induction. The first experiment shows that J48 pruned decision tree algorithm with all attributes is capable of predicting nutritional status as normal, wasted, underweight, and stunted with an accuracy of 92.24%. The second experiment by J48 unpruned decision tree classifier with all attributes also has significant effect on classification and prediction of nutritional status (has accuracy of 92.73%). The third and fourth experiments were designed to evaluate the performance of the Naïve Bayes algorithm in predicting nutritional status and with accuracy performance of 89.68% and 89.84% respectively. The fifth experiment indicates that PART pruned rule induction algorithm with all attributes is highly competent (92.62 % in accurate nutritional status prediction). Finally the sixth experiment, PART unpruned rule induction classifier with all attributes also has promising result on predicting nutritional status which performs 92.82%. In terms of their accuracy performance, all classifiers have good competent performance except Naïve Bayes with all attributes and selected attributes generate 89.68% and 89.84% respectively.

In general, PART pruned rule induction model, J48 unpruned tree and Naives Bayes are appeared with good predictive performance for nutritional status of under-five children. From all the scenarios experimented, all models reveal the better performance in predicting True positive cases or sensitivity; than predictive performance of True negative case or specificity. As sensitivity and specificity has greater importance than general accuracy of the classifier in clinical and medical fields, models are better compared based on WROC area.

The next task in testing the model to decide which one of the six models constitutes a better model/classifier of the 2011 EDHS dataset is evaluated by using ROC area analysis. ROC is the main indicator during algorithm performance selection. In this study, imbalanced data balanced using SMOTE analysis. For this reason, ROC area is the best indicator for SMOTEd data rather than accuracy. With regards to ROC area, a model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicates the stronger evidence for a positive actual state. ROC analysis in which the curve the more to the upper left would indicate a better classifier. Here in this case, the WROC area performance of the algorithms show that PART rule induction, Naïve Bayes with all attributes and J48 pruned decision tree algorithm with all attributes scored the highest area of

97.8% and 97.6% and 97.3% respectively. The lowest WROC keeps account in PART unpruned rule induction with all attributes which is 96%.

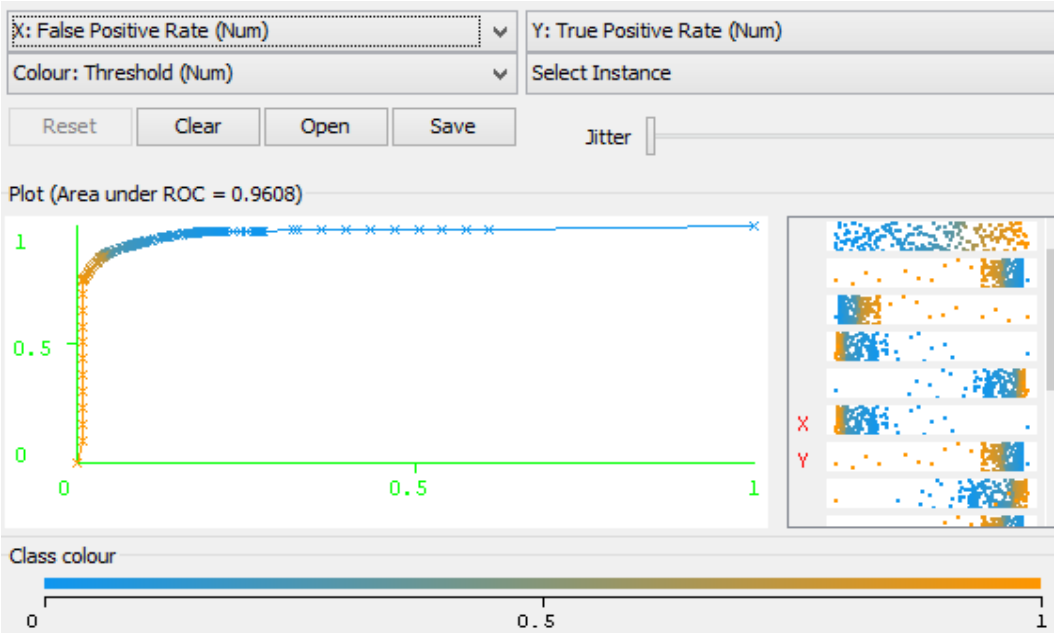

Figure 5.4: Partial ROC area for stunted class

Thus, PART pruned rule induction model with all attributes is selected as the best model and its confusion matrix and ROC area are presented in the next session. The experimental outputs are presented in Annex 3.

5.4.4 Selected model performance and evaluations

During PART pruned rule induction model generation, the effect of the attributes on the model performance was investigated. The full training set containing a total of 14,501 instances were used in all attributes. In addition to the above performance metrics (accuracy, WTPR, WFPR and WROC area) used, relatively PART pruned rule induction with all attributes is more understandable and less complex to human than others model generated. Therefore, the performance of PART pruned rule induction classifier with all attributes gives valuable information in predicting nutritional status as compared to other models.

5.5 Results

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. These three algorithms have literature support (44, 54-55). Models developed using J48 decision tree, Naïve Bayes classifier and PART rule induction algorithms have good accuracy and ROC results during test experiments. Based on these the experiments were designed for four purposes; to investigate the effect of tree pruning methods when building a decision tree model, to observe how attribute selection affects the classification accuracy, to compare J48 decision tree, Naïve Bayes and PART rule induction classifier and to extract significant rules. During test with selected 12 attributes accuracy in J48 and PART rule classifiers showed slight decrease as compared to all 16 attributes results. Due to this Naïve Bayes experiments done using with all and selected attributes.

With regards to effect of pruning, it is obvious that model with grown size of tree make the model difficult to understand and interpret by human as well as generating the rule become challenging. Experiments were done to reduce the complexity of the tree so as to make model more compact and understandable. Therefore, the models are experimented through pruning the tree on the training schemes.

In this study, the model created using PART pruned rule induction classifier registers good performance (i.e 97.8% WROC area) and hence selected for further analysis/rule tracing.

5.6 Rule Extraction

To make decision tree model and PART decision list model more human-readable each path from root to leaf can be transformed into an IF-THEN rule. If the condition is satisfied, the conclusion follows. PART rule induction algorithm is the best known method for deriving rules from classification lists. Both PART rule induction and J48 decision tree classifiers follow decision lists, IF-THEN rule such as PART rule decision list:

CHILDANEM = Not anemic AND
MOTHOCCUP = Working AND
MOTHEDUC = No education AND
TOTAL = >6 AND
SEX = Female AND
MOTHBMI = <18.5: Stunted (17.0/1.0)

The numbers in (parentheses) indicates the number of examples in the leaves. The number of misclassified examples would also be given, in this case 1 (6%) after a slash (/) and hence it is possible to compute the success fraction (ratio) to estimate the level of confidence or likelihood of predictability of the class that tells how much the rule is strong.

PART pruned rule induction model with all attributes produced 346 different rules. However, the researcher selected best interesting rules that cover most of the data points in the study. The other things for selection of rule are new finding and the ability of classifying large instances correctly. The partial PART decision list generated is presented in annex 4.

After the rule extraction, the researcher consulted the literature and domain experts about the generated rules. The discussion was on rules prediction with previous knowledge confirmation. Most of the time one nutritional status predictors (independent variables) could be another predictors later if they applied at the same time. In this study, for example mother's education become predictor for acute and chronic malnutrition of under-five children. For more discussion some of the rules generated by PART pruned rule induction model with all attributes are:

Some specific rules extracted for "Stunted" class

Rule#1: **If** HAZ = <-2SD AND WAZ = -2SD to 2SD: **Then** the class *Stunted (3233.0/8.0)*
This rule revealed 3233(99.7%) correctly out of 3241 instances.

Rule#2: **If** HAZ=<-2SD AND WHZ = -2SD - 2SD AND RESIDENCE = Rural AND CHILDANEM (child anemia level) = Mild AND MOTHEDU = Primary AND TOTAL (Total number of children) = 5-6 AND SIZE (child size at birth) = Normal: **Then** the class *Stunted (12.0/1.0)*

Rule#3: **If** HAZ = <-2SD AND CHILDANEM= Mild AND RESIDENCE = Rural AND MOTHBMI (mothe's body mass index) = <18.5 AND CHILDAGE = 12-23 AND SIZE =Small: **Then** the class *Stunted (12.0/3.0)*

Rule#4: **If** HAZ = <-2SD AND CHILDANEM =Mild AND RESIDENCE = Rural AND WEALTHINDEX = Middle AND MOTHEDUC = No education AND MOTHAGE = 25-29: **Then** the class *Stunted (12.0/3.0)*

Rule#6: **If** HAZ = <-2SD AND CHILDANEM = Moderate AND RESIDENCE = Rural AND MOTHBMI = <18.5 AND EVERHADVAC = Yes AND WEALTHINDEX = Poorest

AND REGION = Affar: **Then** the class *Stunted (17.0/1.0)*

Rule#7: **If** HAZ=<-2SD AND CHILDANEM= Moderate AND MOTHEDUC=No education AND MOTHBMI=18.5-24.9AND MOTHAGE = 25-29 AND TOTAL= 3-4 AND SIZE = Normal AND WEALTHINDEX = Poorest: **Then** the class *Stunted (7.0/2.0)*

The IF-THEN shows rules generated by PART pruned rule induction model to classify nutritional status as stunted (long term or chronic malnutrition). In the case of rule 1, capability of test classification, 3233 (99.7%) correctly and 8 (0.03%) instances classified incorrectly with HAZ attributes (where HAZ <-2SD). HAZ alone matter the fate of the child malnutritioned as stunted keeping other variables constant. In addition to rule1, rule 2 revealed that total number of children in a household, mother's education, mother's body mass index and residence become indictors for child's nutritional status with correct classification of 92.3% instances. Domain area experts and literatures also support these rules (7).

Child age category becomes predictors for stunting with combination of height, i.e, HAZ. In the PART pruned decision list, HAZ would become the first indictor for stunting. Children in rural areas are likely to be stunted and regional variation in the prevalence of stunting in children is substantial. The mother's nutritional status, as measured by her body mass index (BMI), also has a relationship with her child's level of stunting in the case of rules 3 and 6. Relationship is also observed between the household wealth index and the stunting levels of children in rules 6 and 7.

Some specific rules extracted for "Underweight" class

Rule#1**: If** WAZ=<-2SD AND WHZ =-2SD - 2SD AND REGION= Benishangul-Gumuz AND SEX =Female AND MOTHBMI = <18.5: ***THEN*** *the class Underweight (71.0/1.0)*

Rule#2: **If** WAZ=<-2SD AND WHZ = -2SD - 2SD AND REGION = SNNP AND RESIDENCE = Rural AND SIZE = Small AND CHILDAGE = 12-23: ***THEN*** *the class Underweight* ***(18.0)***

Rule#3: **If** WAZ=<-2SD AND WHZ = -2SD - 2SD AND REGION = Benishangul-Gumuz AND CHILDANEM =Moderate AND MOTHEDUC =No education: ***THEN*** *the class Underweight* ***(49.0/3.0)***

Rule#4: **If** WAZ=<-2SD AND WHZ=-2SD-2SD AND CHILDANE=Mild AND RESIDENCE = Rural AND WEALTHINDEX = Poorest AND REGION= SNNP AND MOTHEDUC = No education: ***THEN*** *the class Underweight* ***(30.0)***

Rule#5: **If** WAZ=<-2SD AND CHILDANEM = Not anemic AND CHILDAGE = 36-47 AND MOTHAGE = 25-29 AND MOTHOCCUP = Not working AND SIZE = Small:

THEN *the class Underweight* ***(25.0/1.0)***

The above rules extracted by PART pruned rule induction model applied to predict nutrition status as underweight. The capability of correctly identifying instances is good. In rule 1, when WAZ<-2SD and with others related attributes extracted 71 (98.2%) correctly and 1 (1.8%) incorrectly. Rules reveal that underweight children are experienced in the age groups 12-23 and 36-47 months. This may be explained by the fact that foods for weaning are typically introduced to children in the older age group, thus increasing their exposure to infections and susceptibility to illness. This tendency, coupled with inappropriate or inadequate feeding practices, may contribute to faltering nutritional status among children in these age groups. Being small size at birth has likely to be underweight later in life. According to rule 1, children born to mothers who are thin (BMI less than 18.5) are more likely to be underweight. The proportion of underweight children is higher for those born to uneducated mothers. Rule 4 shows that underweight children decrease as the wealth quintile of the mother increases (52).

Some specific rules extracted for "Wasted" class

Rule#1: **If** WHZ = <-2SD AND WAZ =-2SD-2SD: **Then** the class *Wasted (635.0/3.0)*

Rule#2: **If** WHZ = <-2SD AND CHILDANEM = Mild AND MOTHEDUC = No education: **Then** the class *Wasted (205.0)*

Rule#3: **If** WHZ = <-2SD AND TOTAL = >6 AND MOTHBMI = 18.5-24.9: **Then** the class *Wasted (120.0/2.0)*

Rule#4: **If** WHZ = <-2SD AND HAZ = -2SD- 2SD AND SEX = Male: **Then** the class *Wasted (45.0)*

Rule#5: **If** WHZ = <-2SD AND REGION=Amhara AND SEX=Male AND MOTHBMI=18.5-24.9 AND CHILDANEM=Moderate AND MOTHOCCUP=Not working: **Then** the class *Wasted (7.0)*

Rule#6: **If** WHZ = <-2SD AND REGION=Dire Dawa: **Then** the class *Wasted (101.0/2.0)*

Rule#7: **If** WHZ = <-2SD AND REGION=Affar AND WEALTHINDEX=Poorest AND TOTAL=5-6: **Then** the class *Wasted (85.0/1.0)*

The above decision list rules predicted as wasted (acute malnutrition) by the PART pruned rule induction classifier. Rule 1 revealed that the probability of being wasted is most predicted (99.5%) when WHZ is less than minus two standard deviations keeping all variables constant

(WHZ<-2SD). Rules 4&3 showed that wasting is slightly more likely in male than female children. This showed that currently a child would be diarrheic or/and not breastfeed accordingly. Here, wasting of a child shows not only due to nutritional status but also health status. Children whose mothers have no education or primary education, and in the middle wealth quintiles likely to be wasted. The rule revealed that wealth index of the household is between poorer and rich whose mother has primary education a child become wasted due to not working (jobless or no income source for mother) or lack of health service utilization awareness of mother's.

Some specific rules extracted for "Normal" class

Rule#1: **If** HAZ = -2SD- 2SD AND WAZ = -2SD-2SD AND WHZ = -2SD-2SD AND RESIDENCE = Rural: **Then** the class *Normal (3472.0/1.0)*

Rule#2: **If** HAZ = -2SD- 2SD AND WAZ = -2SD-2SD AND WHZ = -2SD-2SD AND RESIDENCE = Urban AND TOTAL=1-2 AND REGION=Addis Ababa AND SIZE=Normal: **Then** the class *Normal (72.0)*

Rule#3: **If** HAZ = -2SD- 2SD AND WAZ = -2SD-2SD AND WHZ = -2SD-2SD AND RESIDENCE = Urban AND TOTAL=1-2 AND REGION=Addis Ababa AND SIZE=Small AND MOTHOCCUP=Working AND SEX=Male: **Then** the class *Normal (8.0)*

Rule#4: **If** HAZ = -2SD- 2SD AND WAZ = -2SD-2SD AND WHZ = -2SD-2SD AND RESIDENCE = Urban AND TOTAL=1-2 AND REGION=Somali: **Then** the class *Normal (19.0)*

As shown in rule#1, IF-THEN decision list confirmed the three anthropometric measurements, i.e, HAZ, WAZ and WHZ are between minus two and positive two standard deviation (-2SD to 2SD) of those children who have good or normal nutritional status. The first rule gave a correct result for 3472 of the 3473 instances, thus its success fraction is 3472/3473. Rule 1 (when HAZ, WAZ, WHZ=-2SD to 2SD) indicates that the likelihood of a child being at a good nutritional status, keeping all predictors of the class normal is 99.97%.

It is expected that mother's education is likely to improve nutritional status of children through better use of health facilities and better child care practices. Moreover, mother's occupation for her family income source, mother's nutritional status, residence and total number of children

need interventions because of in almost all rules they have important prediction values. If these predictors managed through using different strategy such as giving health education on nutrition, childcare, family planning at community based through rural health extension workers at rural area will improve nutritional status of under-five children.

5.7 Error Rate of the Selected Model

In classification or prediction tasks, the accuracy of the resulting model is measured either in terms of the percentage of instances correctly classified or in terms of "error rate" i.e. the percentage of records classified incorrectly. Classification error rate on pre classified test set is commonly used as an estimate of the expected error rate when classifying new records (13). Errors during each test are averaged to give the average error rate of the model. The classification error rate for the selected model is 7.38%, which means the model has incorrectly classified about 7.38% instances out of their actual classes each time when the model is tested on the test set.

The percentage of incorrectly classified instances indicates the chance with which the developed model misclassifies a new victim out of the actual class. Several reasons may be attributed for increased error rate from the models. First, algorithms differ in their capability as observed from comparisons of performance measures. Second, attributes may not be included in the collection and study might have influenced it. All the models of the predictive performance in identifying True Positive cases of model are higher than identifying True Negative cases. Consequently, the model tends to misclassify instances to some other classes.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as healthcare sector, has been applied for patient survival analysis, prediction of diagnosis, for outcomes measurement, to improve patient care and decision-making etc. However, the potentials of data mining have not yet been used in predicting nutritional status of under-five children in Ethiopia. In this study, the objective was to design a predictive model for nutritional status of under-five children using data mining techniques using 2011 EDHS dataset. The model would be used in the future so as to help policy makers and health care providers in the country to identify children who are at risk. Furthermore, such a predictive model might be applied in assisting under-five malnutrition prevention and control activities in the country.

The hybrid, iterative methodology, was employed in this study which consists of six basic steps such as problem domain understanding, data understanding, data preparation, data mining, and evaluation of the discovered knowledge and use of the discovered knowledge.

In order to generate interesting rule from the huge data collected in the 2011 EDHS, a total of 9607 instances and 17 attributes were applied. Knowledge discovery in dataset was employed after SMOTE technique was applied which is an automatic operation where minority classes are over sampled to make the target attribute balance.

In this particular research, independent variables/attributes were: mother's age, region, residence, mother's occupation, mother's education, mother's BMI, wealth index, total number of children, age of child, sex of child, ever had vaccination, size of child at birth, child anemia level, HAZ, WAZ, WHZ and dependent variable/attributes was under nutrition status.

The findings clearly suggested that most of the above attributes had strong relation with nutritional status of under-five children in the demographic and health survey data. All the

selected attributes were used in the analysis using J48 decision tree, Naïve Bayes and PART rule induction algorithms.

Several models were built during experimentation that could predict the risk of under-five malnutrition. Among these models, PART pruned rule induction model with all attributes showed an interesting predictive accuracy result of 92.6% and ROC area of 97.8%.

In summary, region, residence, mother's education, wealth index of the household, child age are major contributor of under nutrition and predictive model is developed using PART pruned rule induction with all attributes to predict nutritional status of under-five children in Ethiopia.

6.2 Recommendations

In this study, an attempt was made to explore the 2011 EDHS dataset and to provide an initial insight into the potential applicability of data mining techniques in predicting nutritional status of under-five children based on demographic, health and socioeconomic characteristics. Reducing child malnutrition and improving child health status, through appropriate interventions, requires better understanding of the main demographic, health and socioeconomic determinants. Thus, based on the result of the research learned by PART pruned rule induction algorithm, the following recommendations were made by the researcher.

- Models presented in this study showed that under nutrition can be reduced substantially by intervening in certain socio-economic and demographic factors so that probability of under-five malnutrition can be minimized. Thus models can be used in formulating child nutrition programs and child health policies.
- This study considered demographic and health survey dataset to predict nutritional status of under-five children. So that future studies might need to discover knowledge and patterns in other domain areas such as clinical datasets.
- It has been observed that developing many other classifiers for prediction with the short period of time given to this research was unlikely. Therefore, to enhance the performance of the present model further research in nutritional status of under-five area incrementally should be done using many more mining techniques to improve the predictive model accuracy.
- This research developed predictive model for under five children nutritional status prediction. However, there is a need for knowledge based system for predicting under five children nutritional status. This should be future research direction.
- Organizations working on under-five children nutrition should also work on the determinant identified findings.

REFERENCES

1. UNESCO Report, The Review of Health and Nutrition indicators in early childhood, March, 2012.
2. Ethiopia Federal Ministry of Health, Protocol for the management of severe acute malnutrition, Ethiopia, Addis Ababa, March 2007.
3. Tackling child malnutrition, Save the Children, February 2012.
4. Aweke K, Habtamu F and G Akalu, *Nutritional status of children in food insecure households in two districts of north showa zone*, Ethiopia, Volume 12 No. 2 April 2012.
5. MDG Report 2012, Assessing Progress in Africa toward the Millennium Development Goals, September 2012
6. Teklebrhan Tema Beyene, *Predictors of Nutritional Status of Children Visiting Health Facilities in Jimma Zone*, South West Ethiopia, Research article, Department of Nursing, Jimma University, Ethiopia, 16 September 2012.
7. Central Statistical Agency, Addis Ababa: Ethiopia Demographic and Health Survey 2011, Ethiopia ICF International Calverton, Maryland, USA March 2012.
8. Ethiopian Demographic Health Survey, Central Statistics authority of Ethiopia and Macro International Inc. 2005.
9. Solomon Amsalu and Zemene Tigabu, Risk factors for severe acute malnutrition in children under the age of five: A case-control study, Original article, Ethiop.J.Health Dev. 2008.
10. Beka Teshome, Wambui Kogi-Makau, Zewditu Getahun and Girum Taye. Magnitude and determinants of stunting in children under five years of age in food surplus region of Ethiopia: The case of West Gojam Zone, Ethiop. J. Health Dev. 2009.
11. Alemu Mekonnen, Bekele Tefera, Tassew Woldehanna, Nicola Jones, John Seager, Tekie Alemu and Getachew Asgedom. Child nutritional status in poor Ethiopian households: The role of gender, assets and location. Working paper No.26, Young Lives, Save the Children UK, 2005.
12. Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W, Kurgan Lukasz A. Data Mining: A Knowledge Discovery Approach. New York, USA: Springer Science Business Media LLC; 2007.
13. Berry Michael J.A., Linoff Gordon S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Second Edition. Wiley Publishing, Inc., United States

of America. 2004.

14. Laxman, S. & Sastry, P. S., A Survey of Temporal Data Mining. Sâdhanâ, 31(2):173–198, 2006.
15. Bath Peter A. Data Mining in Health and Medical Information: Annual review of Information Science and Technology. Blaise Cronin, editor; Vol 38. USA; Information Today Inc; 2004.
16. Larose Daniel T. Discovering Knowledge in Data - An Introduction to Data Mining. New Jersey, USA: John Wiley & Sons Inc; 2005.
17. Hand, D., Mannila, H. & Smyth. Principles of data mining. Massachusetts. Massachusetts Institute of Technology press, 2001.
18. Witten, Ian. H. & Frank, Eibe. Data mining: practical machine learning tools and techniques. Second edition. San Francisco: Morgan Kaufmann Publishers, 2005.
19. Kurgan, L. A. & Musilek, P. A survey of knowledge discovery and data mining process models: The Knowledge Engineering Review, 21(1):1-24, 2006.
20. Refaat, M.. Data preparation for data mining using SAS. San Francisco; Morgan Kaufmann Publishers, 2007.
21. Azevedo, A. & Santos, M.F. KDD, SEMMA and CRISP-DM: A parallel overview. IADIS European conference data mining: 182-185, 2008.
22. Fayyad,U., Piatetsky-Shapiro,G., and Smyth, P. From data mining to knowledge discovery in databases. American Association for Artificial Intelligence: 36-51, 1996.
23. Brachman, R. J. & Anand, T. The process of knowledge discovery in databases. 1996.
24. Statistical Analiysis Software (SAS), Retrieved January 02, 2013, from http://www.sas.com/offices/europe/uk/technologies/analytics datamining/miner/semma.html.
25. Ana Azevedo, Manuel Filipe S. Data Mining Standards, Knowledge Discovery in Databases: Data Mining; 2008.
26. Glover, S., Rivers,P., Asoh,D., Piper,C. and Murph,K. Data mining for health executive decision support: an imperative with a daunting future! Health Services Management Research, 23(1),42-44, 2010.
27. Bramer Max. Principles of Data Mining. London. Springer-Verlag Limited; 2007.
28. Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery. 3rd ed.

Two Crows Corporation. 500 Falls Road, Potomac, USA, 2005.

29. Supervised and Unsupervised learning: From http://sisla06.samsi.info/jpal/mult1031.pdf. Retrieved on January 26, 2013.
30. Han Jiawei, Kamber Micheline. Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann Publishers; 2001.
31. Mehamed Kantardzic J.B. Data Mining-Concepts, Models, Methods, and Algorithms. USA:John Wiley & Sons Publication Inc; 2003.
32. Velickov, S. and Solomatine, D. Predictive Data Mining: Practical Example, Moscow, Russia, 2000.
33. Gerritsen, R. Assessing Loan Risks: A Data Mining. Case Study, 1999.
34. Han, J & Kamber, M., Data mining: concepts and techniques. (2nd ed.). San Francisco: Morgan Kaufmann Publishers, 2006.
35. Feyen HaLP. Data Mining and Strategic Marketing in the Airline Industry. Online.[Access date January 10,2013.
36. Bouckaert Remco R., Frank Eibe, Hall Mark, Richard Kirkby, Reutemann Peter, Seewald Alex, Scuse David. WEKA Manual for Version 3-6-2. University of Waikato, Hamilton, New Zealand. January 11, 2010.
37. David L. Olson and Dursun D. Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg, 2008.
38. Weiss Sholom M., Zhang Tong. Performance Analysis and Evaluation. In:Ye Nong, Editor. The Hand Book of Data Mining. New Jeresy. USA: Lawerence Erlbaum Associates Inc; 2003.
39. Ifeachor C E, Hamadicharef B. Receiver Operating Curve Analysis in The Evaluation of Intelligent Medical Systems. UK: University of Plymouth. Drake Circus Plymouth PL4 8AA, Devon; 2004.
40. Melanie C. Page, Sanford L Braver, David P. MacKinnon. Levine's Guide to SPSS for Analysis of Variance. London: Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey; 2003.
41. Sellappan Palaniappan, Awang Rafiah. *Intelligent Heart Disease Prediction System Using Data Mining Techniques*. International Journal of Computer Science and Network Security. 343-350 Aug 8, 2008.
42. Philip Baylis., Better health care with data mining. SPSS Inc. WPDMHC-0699, 1999

43. S. P. Deshpande and V. M. Thakare. *Data Mining System and Applications*: A Review. International Journal of Distributed and Parallel systems, Volume 1, Number 1: 445-463, 2010.
44. Tesfahun Hailemariam, Application of Data Mining Techniques to Predict Adult Mortality: The Case of BRHP, Master's Thesis, AAU, Ethiopia; pp 39-85 June 2012.
45. Kolar HR. Caring for Healthcare. Health Management Technology; 22(4): 48-47, 2001.
46. Tylor P. From Patient Data to Medical Knowledge: The Principles and Practices of Health Informatics. UK: Backwell Publishings ltd; 2008.
47. Kincade K. Data Mining: Digging for Healthcare Gold. Insurance & Technology, 23(2): 2-7, 1998.
48. Soni Jyoti, Ansari Ujma, Sharma Dipesh, Soni Sunita. *Predictive Data Mining for Medical Diagnosis*: An Overview of Heart Disease Prediction. International Journal of Computer Applications, India. 17(8): 0975 – 8887, March 2011.
49. Xu Dezhi and Gamage Upeksha , Rule Based Classification to Detect Malnutrition in Children , School of Information Science & Engineering, Central South University, Changsha, China, Vol. 3 No. 1, pp 428, Jan 2011.
50. Retrived on Februar, 16, 2013 http://shodhganga.inflibnet.ac.in/bitstream/10603/2588/9/09_chapter%202.pdf
51. Andrew Ryan Nutritional Status of Children at Rise2Shine Facility - Rise2Shine Fond Parisien, Haiti, pp 1-4, March 19, 2012.
52. Abdu Kedir Seid. Health and Nutritional Status of Children in Ethiopia: Do Maternal Charateristics Mattere? Copenhagen University, Denmark; pp 1-18. August, 2012.
53. Shegaw A. Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project. M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia; 2002.
54. Be'emnetu Tekabe. Predicting the patterns of Under-five Mortality Using Data Mining Technology. The Case of Butajira Rural Health Program. Master's Thesis; Addis Ababa University, Addis Ababa, Ethiopia; pp 8-10 and 104-106, June 2002.
55. Biset Desalegn. Predicting Low Birth Weight Using Data Mining Techniques on Ethiopia Demographic and Health Survey Datasets. Master's Thesis: AAU, Ethiopia; pp 4-10 & 73-75; June 2011.

56. Chakrabarti .S ,Earl C., Eibe F., Ralf H.G., Jaiwei H. , Xia J., Micheline K., Sam S. L.,Thomas P. ,Richard E. ,Dorian P., Mamdouh R.,Markus S.,Toby J. and Witten H. *Data mining know it all*. Morgan Kaufmann Publishers 30 Corporate Drive, Suite 400 Burlington, United States. 2009.
57. Ping-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data mining*. Pearson Educ. Inc. 2006.
58. Xiaohua Hu. *DB-HReduction: A Data Preprocessing Algorithm for Data Mining Applications.* College of Information Science and Technology, Drexel University Philadelphia, PA 19104, U.S.A. 2003.
59. Colin Shearer. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Ware Housing Volume 5, Number 4. 2000.
60. Nitesh V, Chawla, Kevin W. Bowyer, Lawrence, O.Hall, W. and Philip K.SMOTE: Synthetic Minority Over Sampling Technique. Department of Computer Science and Engineering, ENB 188. University of South Florida; 2002.

APPENDIXES

ANNEX 1: Description of the Selected Attributes

No	Field name	Descriptions	Corresponding values
1	MOTHAGE	Mother's age group	15-19,20-24,25-29,30-34,35-39,40-44,45-49
2	REGION	Regions	Tigray, Affar, Amhara,Oromiya,Somali, Benishangul-Gumuz, SNNP, Gambela,Harari,Addis Ababa, Dire Dawa
3	RESIDENCE	Living place	Urban ,Rural
4	MOTHEDUC	Mother education level	No education, Primary, Secondary, Higher
5	WEALTHINDEX	Wealth index/ quartiles	Poorest, Poorer, Middle, Richer, Richest
6	MOTHBMI	Mother Body Mass Index	Thin=<18.5, Normal=18.5-24.9, Over/Obse=>=25
7	MOTHOCCUP	Mother occupation status	Working, Not working
8	TOTAL	Total number of children / Parity	1-2, 3-4, 5-6, >6 children
9	SEX	Sex of child	Female, Male
10	CHILDAGE	Child age group/category	<6, 6-11, 12-23, 24-35, 36-47,48-59 years
11	SIZE	Size of child at birth	Small (<2.5kg), Normal (2.5-4kg), Large (>4kg)
12	CHILDANEM	Child anemia level	Not anemic, Mild, Moderate, Severe
13	EVERHADVAC	Child ever had vaccinated	Yes, No
14	HAZ	Height for age Z-score	<-2SD, -2SD-2SD, >2SD (Standard deviations)
15	WAZ	Weight for age Z-score	<-2SD, -2SD-2SD, >2SD (Standard deviations)
16	WHZ	Weight for height Z-score	<-2SD, -2SD-2SD, >2SD (Standard deviations)
17	NUTSTATUS	Nutritional Status of the child	Normal, Wasted, Underweight or Stunted

ANNEX 2: J48 pruned decision tree before SMOTE with all attributes

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: NutStatusOfU5InEthiopia-weka.filters.unsupervised.attribute.ReplaceMissingValues
Instances: 9607
Attributes: 17
MOTHAGE, REGION, RESIDENCE, MOTHEDUC, WEALTHINDEX,
MOTHBMI, MOTHOCCUP, TOTAL, SEX, CHILDAGE, SIZE, CHILDANEM
EVERHADVAC, HAZ, WAZ, WHZ, NUTSTATUS
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===

```
WHZ = -2SD-2SD
|   HAZ = <-2SD
|   |   WAZ = -2SD-2SD: Stunted (1612.0/8.0)
|   |   WAZ = <-2SD: Underweight (1940.0/645.0)
|   |   WAZ = >2SD: Stunted (0.0)
|   HAZ = -2SD- 2SD
|   |   WAZ = -2SD-2SD: Normal (4428.0/6.0)
|   |   WAZ = <-2SD: Underweight (133.0/1.0)
|   |   WAZ = >2SD: Normal (4.0)
|   HAZ = >2SD: Normal (198.0)
WHZ = <-2SD: Wasted (1128.0/25.0)
WHZ = >2SD
|   HAZ = <-2SD: Stunted (81.0)
|   HAZ = -2SD- 2SD: Normal (83.0/5.0)
|   HAZ = >2SD: Stunted (0.0)

Number of Leaves  :   11
Size of the tree :    16
Time taken to build model: 5.59 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        8899               92.6304 %
Incorrectly Classified Instances       708                7.3696 %
Kappa statistic                          0.8899
Mean absolute error                      0.0492
Root mean squared error                  0.1602
Relative absolute error                 14.8034 %
Root relative squared error             39.3214 %
Total Number of Instances             9607
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.734	0.008	0.968	0.734	0.835	0.966	Stunted
	0.951	0.075	0.691	0.951	0.801	0.955	Underweight
	0.998	0.002	0.998	0.998	0.998	0.998	Normal
	0.997	0.003	0.978	0.997	0.987	0.997	Wasted
Weighted Avg.	0.926	0.015	0.942	0.926	0.927	0.984	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
1716	612	10	1	a = Stunted
49	1378	1	21	b = Underweight
8	0	4702	3	c = Normal
0	3	0	1103	d = Wasted

ANNEX 3: Summary of the Output of the Classifiers

Ex#1:J48 pruned decision tree with all attributes

Test mode:10-fold cross-validation
Classifier model (full training set)
Number of Leaves: 590
Size of the tree: 808
Stratified cross-validation

Correctly Classified Instances	13376	92.2419 %
Incorrectly Classified Instances	1125	7.7581 %
Total Number of Instances	14501	

Attributes: 17

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.861	0.043	0.904	0.861	0.882	0.954	Stunted
	0.841	0.054	0.794	0.841	0.817	0.942	Underweight
	0.998	0.002	0.996	0.998	0.997	0.998	Normal
	0.999	0.004	0.98	0.999	0.989	0.997	Wasted
Weighted Avg.	0.922	0.026	0.923	0.922	0.923	0.973	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
4027	631	19	1	a = Stunted
417	2438	1	42	b = Underweight
8	0	4702	3	c = Normal
2	1	0	2209	d = Wasted

Ex#2: J48 unpruned decision tree with all attributes

Test mode: 10-fold cross-validation
Classifier model (full training set)
Number of Leaves: 1491
Size of the tree: 983
Stratified cross-validation
Correctly Classified Instances 13447 92.7315 %
Incorrectly Classified Instances 1054 7.2685 %
Total Number of Instances 14501
Attributes: 17
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.899	0.052	0.891	0.899	0.895	0.941	Stunted
	0.822	0.042	0.831	0.822	0.826	0.933	Underweight
	0.994	0.002	0.996	0.994	0.995	0.998	Normal
	0.982	0.003	0.985	0.982	0.983	0.994	Wasted
Weighted Avg.	0.927	0.026	0.927	0.927	0.927	0.966	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
4206	452	18	2	a = Stunted
486	2383	1	28	b = Underweight
24	0	4686	3	c = Normal
4	34	2	2172	d = Wasted

Ex#3: J48 pruned decision tree with 12 selected attributes

Test mode: 10-fold cross-validation
Classifier model (full training set
Number of Leaves: 11
Size of the tree: 16
Correctly Classified Instances 13135 90.58 %
Incorrectly Classified Instances 1366 9.42 %
Total Number of Instances 14501
Attributes: 13
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.72	0.001	0.998	0.72	0.836	0.96	Stunted
	0.985	0.111	0.688	0.985	0.811	0.946	Underweight
	0.998	0.002	0.996	0.998	0.997	0.998	Normal
	0.999	0.004	0.98	0.999	0.989	0.997	Wasted
Weighted Avg.	0.906	0.024	0.932	0.906	0.907	0.975	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
3369	1289	19	1	a = Stunted
0	2855	1	42	b = Underweight
8	0	4702	3	c = Normal
0	3	0	2209	d = Wasted

Ex#4: J48 unpruned decision tree with 12 selected attributes

Test mode: 10-fold cross-validation
Classifier model (full training set)
Number of Leaves: 1491
Size of the tree: 983
Stratified cross-validation
Correctly Classified Instances 13295 91.6833 %
Incorrectly Classified Instances 1206 8.3167 %
Total Number of Instances 14501
Attributes: 13
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.868	0.053	0.885	0.868	0.876	0.954	Stunted
	0.817	0.054	0.791	0.817	0.804	0.948	Underweight
	0.994	0.002	0.996	0.994	0.995	0.997	Normal
	0.986	0.003	0.984	0.986	0.985	0.995	Wasted
Weighted Avg.	0.917	0.029	0.917	0.917	0.917	0.973	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
4059	599	18	2	a = Stunted
498	2369	1	30	b = Underweight
24	0	4686	3	c = Normal
3	27	1	2181	d = Wasted

Ex#5: Naive Bayes Classifier with all selected attributes

Test mode:10-fold cross-validation
Classifier model (full training set)
Stratified cross-validation
Correctly Classified Instances 13005 89.6835 %
Incorrectly Classified Instances 1496 10.3165 %
Total Number of Instances 14501
Attributes: 17
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.736	0.021	0.943	0.736	0.827	0.961	Stunted
	0.916	0.105	0.685	0.916	0.784	0.949	Underweight
	0.998	0.002	0.995	0.998	0.996	0.998	Normal
	0.997	0.004	0.98	0.997	0.988	0.998	Wasted
Weighted Avg.	0.897	0.029	0.914	0.897	0.898	0.976	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
3443	1215	19	1	a = Stunted
201	2654	1	42	b = Underweight
8	0	4702	3	c = Normal
0	3	3	2206	d = Wasted

Ex#6: Naive Bayes Classifier with selected attributes

Test mode:10-fold cross-validation
Classifier model (full training set)
Stratified cross-validation

Correctly Classified Instances	13028	89.8421 %
Incorrectly Classified Instances	1473	10.1579 %
Total Number of Instances	14501	

Attributes: 13

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.736	0.019	0.948	0.736	0.829	0.961	Stunted
	0.923	0.105	0.687	0.923	0.788	0.948	Underweight
	0.998	0.002	0.995	0.998	0.997	0.998	Normal
	0.998	0.004	0.98	0.998	0.989	0.998	Wasted
Weighted Avg.	0.898	0.028	0.916	0.898	0.9	0.976	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
3444	1214	19	1	a = Stunted
180	2675	1	42	b = Underweight
8	0	4702	3	c = Normal
0	3	2	2207	d = Wasted

Ex#7: PART unpruned rule induction classifier with all attributes

== Run information ===
Test mode: 10-fold cross-validation
Classifier model (full training set)
Number of Rules: 346
Stratified cross-validation

Correctly Classified Instances	13432	92.6281 %
Incorrectly Classified Instances	1069	7.3719 %
Total Number of Instances	14501	

Attributes: 17

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.875	0.044	0.905	0.875	0.89	0.961	Stunted
	0.843	0.05	0.808	0.843	0.825	0.955	Underweight
	0.997	0.002	0.996	0.997	0.997	0.999	Normal
	0.992	0.003	0.983	0.992	0.987	0.998	Wasted
Weighted Avg.	0.926	0.025	0.927	0.926	0.926	0.978	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
4093	565	19	1	a = Stunted
419	2443	1	35	b = Underweight
9	0	4701	3	c = Normal
2	15	0	2195	d = Wasted

Ex#8: PART unpruned rule induction classifier with all attributes

Test mode: 10-fold cross-validation
Classifier model (full training set)
Number of Rules: 814
Stratified cross-validation
Correctly Classified Instances 13460 92.8212 %
Incorrectly Classified Instances 1041 7.1788 %
Total Number of Instances 14501
Attributes: 17
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.908	0.055	0.888	0.908	0.898	0.938	Stunted
	0.816	0.038	0.842	0.816	0.829	0.912	Underweight
	0.992	0.003	0.994	0.992	0.993	0.996	Normal
	0.981	0.003	0.985	0.981	0.983	0.992	Wasted
Weighted Avg.	0.928	0.027	0.928	0.928	0.928	0.96	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
4249	405	23	1	a = Stunted
503	2366	1	28	b = Underweight
32	2	4676	3	c = Normal
3	36	4	2169	d = Wasted

Ex#9: PART pruned rule induction classifier with 12 selected attributes

Test mode: 10-fold cross-validation
Classifier model (full training set)
Number of Rules: 269
Stratified cross-validation
Correctly Classified Instances 13301 91.7247 %
Incorrectly Classified Instances 1200 8.2753 %
Total Number of Instances 14501
Attributes: 13
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.848	0.045	0.9	0.848	0.873	0.966	Stunted
	0.837	0.06	0.777	0.837	0.806	0.959	Underweight
	0.998	0.002	0.996	0.998	0.997	0.988	Normal
	0.997	0.003	0.981	0.997	0.989	0.988	Wasted
Weighted Avg.	0.917	0.028	0.919	0.917	0.918	0.976	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
3966	692	19	1	a = Stunted
432	2427	1	38	b = Underweight
8	0	4702	3	c = Normal
0	6	0	2206	d = Wasted

Ex#10: PART unpruned rule induction classifier with 12 selected attributes

Test mode:10-fold cross-validation
Classifier model (full training set)
Stratified cross-validation
Number of Rules : 868
Correctly Classified Instances 13174 90.8489 %
Incorrectly Classified Instances 1327 9.1511 %
Total Number of Instances 14501
Attributes: 13
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.877	0.068	0.86	0.877	0.868	0.954	Stunted
	0.775	0.052	0.789	0.775	0.782	0.941	Underweight
	0.991	0.003	0.994	0.991	0.993	0.997	Normal
	0.976	0.003	0.984	0.976	0.98	0.994	Wasted
Weighted Avg.	0.908	0.033	0.909	0.908	0.908	0.971	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
4101	558	18	1	a = Stunted
623	2245	1	29	b = Underweight
38	1	4670	4	c = Normal
4	43	7	2158	d = Wasted

ANNEX 4: Partial PART decision list generated output for the selected Model

=== Run information ===
Scheme:weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation: NutStatusOfU5InEthiopia-
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C1-K5-P100.0-S1-
weka.filters.supervised.instance.SMOTE-C2-K5-P100.0-S1
Instances: 14501
Attributes: 17
MOTHAGE, REGION, RESIDENCE, MOTHEDUC, WEALTHINDEX, MOTHBMI, MOTHOCCUP, TOTAL, SEX, CHILDAGE, SIZE, CHILDANEM, EVERHADVAC, HAZ, WAZ, WHZ, NUTSTATUS
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===

PART decision list

WHZ = <-2SD AND
WAZ = -2SD-2SD: Wasted (635.0/3.0)

HAZ = -2SD- 2SD AND
WAZ = -2SD-2SD AND
WHZ = -2SD-2SD AND
RESIDENCE = Rural: Normal (3472.0/1.0)

HAZ = <-2SD AND
CHILDANEM = Mild AND
RESIDENCE = Rural AND
REGION = Benishangul-Gumuz AND
MOTHEDUC = No education AND
EVERHADVAC = Yes AND
SEX = Male: Stunted (10.0)

CHILDANEM = Not anemic AND
CHILDAGE = 36-47 AND
MOTHAGE = 25-29 AND
MOTHOCCUP = Not working AND
SIZE = Small: Underweight (25.0/1.0)

CHILDANEM = Not anemic AND
MOTHEDUC = No education AND
SIZE = Normal AND
TOTAL = >6 AND
WEALTHINDEX = Poorest AND
MOTHAGE = 30-34: Stunted (14.0)

CHILDANEM = Not anemic AND
MOTHOCCUP = Working AND
MOTHEDUC = No education AND
TOTAL = >6 AND
SEX = Female AND
MOTHBMI = <18.5: Stunted (17.0/1.0)

CHILDANEM = Not anemic AND
MOTHEDUC = No education AND
CHILDAGE = 12-23 AND
MOTHBMI = <18.5: Underweight (29.0/4.0)

CHILDANEM = Not anemic AND
MOTHEDUC = No education AND
MOTHAGE = 35-39 AND
SIZE = Normal AND
CHILDAGE = 36-47: Underweight (35.0/2.0)

CHILDANEM = Not anemic AND
MOTHEDUC = No education AND
SIZE = Large AND
SEX = Male AND
REGION = Benishangul-Gumuz AND
EVERHADVAC = Yes: Stunted (23.0/2.0)

WAZ=<-2SD AND HAZ = <-2SD AND
CHILDANEM = Moderate AND
MOTHBMI = 18.5-24.9 AND
SEX = Male AND
MOTHOCCUP = Not working AND
MOTHAGE = 25-29: Underweight (40.0/9.0)

HAZ = <-2SD AND
CHILDANEM = Moderate AND
RESIDENCE = Rural AND
MOTHBMI = <18.5 AND
EVERHADVAC = Yes AND
WEALTHINDEX = Poorest AND
REGION = Affar: Stunted (17.0/1.0)

HAZ = <-2SD AND
CHILDANEM = Moderate AND
RESIDENCE = Rural AND
MOTHBMI = <18.5 AND
SEX = Male AND
CHILDAGE = 24-35: Stunted (14.0/2.0)

WAZ=<-2SD AND HAZ = <-2SD AND
CHILDANEM = Moderate AND
SIZE = Normal AND
MOTHAGE = 30-34 AND
MOTHBMI = <18.5 AND
TOTAL = 5-6: Underweight (17.0)

WAZ=<-2SD AND HAZ = <-2SD AND
CHILDANEM = Mild AND
WEALTHINDEX = Poorest AND
REGION = Tigray AND
SIZE = Normal: Underweight (19.0)

WAZ=<-2SD AND WHZ = -2SD-2SD
AND RESIDENCE = Rural AND
CHILDANEM = Mild AND
MOTHEDUC = Primary AND
SEX = Male AND
TOTAL = 3-4: Underweight (20.0/1.0)

WAZ=<-2SD AND WHZ = -2SD-2SD
AND CHILDANEM = Severe AND
WEALTHINDEX = Poorest AND
MOTHBMI = 18.5-24.9 AND
TOTAL = 5-6: Underweight (33.0/1.0)

WHZ = <-2SD AND
CHILDANEM = Mild AND
MOTHEDUC = No education: Wasted (205.0)

WHZ = <-2SD AND
REGION = Affar AND
WEALTHINDEX = Poorest: Wasted (207.0/3.0)

HAZ = -2SD- 2SD AND
WAZ = -2SD-2SD AND
WHZ = -2SD-2SD AND
RESIDENCE = Rural: Normal (3472.0/1.0)

WHZ = <-2SD AND
MOTHAGE = 25-29 AND
TOTAL = 3-4 AND
HAZ = -2SD- 2SD: Wasted (69.0/1.0)

WHZ = <-2SD AND
TOTAL = >6 AND
MOTHBMI = 18.5-24.9: Wasted (120.0/2.0)