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

DEPARTMENT OF HEALTH INFORMATICS

PREDICTING THE OCCURRENCE OF UNDER NUTRITION STATUS OF UNDER- FIVE CHILDREN IN ETHIOPIA

BY

CHALUMA KUMELA MENGESHA

JUNE, 2015

ADDIS ABABA, ETHIOPIA

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

DEPARTMENT OF HEALTH INFORMATICS

PREDICTING THE OCCURRENCE OF UNDER NUTRITION STATUS OF UNDER- FIVE CHILDREN IN ETHIOPIA

BY

CHALUMA KUMELA MENGESHA

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

JUNE, 2015

ADDIS ABABA, ETHIOPIA

Affiliation:	Addis Ababa University
Programme:	Msc in Health Informatics
Project Title:	Predicting the occurrence of under nutrition status of under- five children in Ethiopia
Student Name:	Chaluma Kumela Mengasha
Date:	June, 2015

Approval

Name	Title	Signature	Date
Dr. Gashaw Kebede	Advisor	_____	_____
Dr. Abera Kumie	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

Declaration

I declare and confirm that this project is my original work. I have followed all ethical principles. All references that are included in this project have been given recognition through the citation. I confirm that I have cited and referenced all sources used in this document.

This project work is submitted in partial fulfillment of the requirements for the Master of Science degree from the school of Graduate Studies at Addis Ababa University. I seriously declare that this project work has not been submitted to any other institution anywhere for the award of any academic degree.

Name of the student

Chaluma Kumela Mengasha: Signature_____ Date_____

Advisors

Gashaw Kebede PhD: Signature_____ Date_____

Abera Kumie PhD: Signature_____ Date_____

Acknowledgements

First of all, I gratefully express my deepest gratitude to the almighty God for his guidance, help, supports and because he let me see new success in my life- Glory to God.

I heartily, thank my advisors, Dr. Gashaw Kebede and Dr. Abera Kumie for their encouragement, guidance, constructive comments, support and their help that enabled me to understand and develop this project work.

I forward my sincere gratitude to Central Statistics Agency staff for providing the data set used in this project work.

I also acknowledge Federal Ministry of Health staffs- especially; Ato Ayele Birara, Ato Agegnaw Aweke and organized team for their help in making me understand the depth of under-nutrition problem and selecting extracted rules.

I would like to thank Addis Ababa University, School of Information Science and School of Public Health in financial support and over all facilitation of this project from the beginning until the end.

Also I thank my class mates Dereje Ayelow, Fikru Negessa and Tarekegn Tedessa for sharing their idea, support and comments during execution of this project work.

It is difficult to mention all individuals and appreciate the contributions every one gave me. I hope to say my heartfelt thank to all those who have helped me when I carry out this project work.

Table of Contents

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

List of Tables

List of Figures

Abbreviation

ANC: Anti Natal Care

ARFF: Attribute Related File Format

CRISP-DM: Cross- Industry Standard Process for Data Mining

CSA: Central Statistics Agency

CLTS: Community-Led Total Sanitation

DM: Data Mining

EDHS: Ethiopian Demographic and Health Survey

EMDHS: Ethiopian Mini Demographic and Health Survey

EOS: Expended Outreach Strategy

GDP: Growth Domestic Product

HSDP-IV: Health Sector Development Programme IV

KDD: Knowledge Discovery from Database

KDP: Knowledge Discovery Process

SD: Standard Deviation

FMoH: Federal Ministry of Health

UNICEF: United Nations International Children's Fund

WHO: World Health Organization

WEKA: Waikato Environment for Knowledge Analysis.

MUAC: Mid Upper Arm Circumstances.

HAZ: Height-for- age Z-score

WHZ: Weight-for- height Z- score

WAZ: Weight-for- age Z- score

PFSA: Pharmaceutical Funding and Supply Agency

RUTF: Read to Use Therapeutic Food

NNP: National Nutrition Programme

ROC: Receiver Operating Characteristics Analysis

TPR: True Positive Rate

FPR: False Positive Rate

WASH: Water, Sanitation and Hygiene

Summary

Background: Under-nutrition mostly affects under-five children in Ethiopia. To overcome this problem, Ethiopian government has formulated and implemented various programs since long period of time. Thus, this project is intended to build a predictive model that supports the designing of intervention strategies.

Objective: To build predictive model that predicts the occurrence of under- nutrition among under- five children in Ethiopia.

Methodology: This project work followed CRISP-DM Methodology of Knowledge Discovery Process. This project used secondary data from 2014 EMDHS children's nutrition database of CSA. WEKA 3.7.10 version and data mining techniques such as J48 decision tree, Naïve Bayes and PART rule induction classifiers have been applied.

Result: For building the model that predicts under-five nutritional status; J48 was selected as the best performed algorithms compared to Naïve Bayes and PART rule induction by conducting 15 experimentations. The result of the experiments showed J48 was performing with accuracy, WTPR and WROC of 63.5%, 63.6% and 82.2% respectively. This algorithm generated 173 rules out of which only 11 rules were selected.

Conclusion: The potential of data mining application have not yet been used on 2014 EMDHS data for predicting occurrence of under-nutrition among under-five in Ethiopian. Thus, by this project work, algorithm of data mining (J48) was applied on child nutritional status of 2014 EMDHS data .Predictive models (11 rules) that help to predict child under-nutrition at national.

Recommendation: FMoH in collaboration with regional health bureaus should strengthen health care services such as ANC visits, institutional delivery, and mothers' education, improving drinking water, Hygiene and sanitation as well as focusing and giving special care for children whose age between 36 and 47months in order to manage under-nutrition problems.

CHAPTER ONE

1. Introduction

1.1 Background

Malnutrition is an important public health issue particularly for children under five years old who have a significantly higher risk of mortality and morbidity(1). It can be categorized into under-nutrition and over-nutrition. Child under-nutrition is highly prevalent in low and middle income countries including Ethiopia. It is the most important risk factor for the burden of diseases that causing about 300,000 deaths per year directly and indirectly (2).

Child under-nutrition involves complex processes at multiple levels; from individual to the household, to the community, to the national and international levels. Under-nutrition commonly affects all groups in community, but infants and young children are the most vulnerable because of their high nutritional requirements for growth and development (4).

Anthropometric measurement is the most commonly used method to assess child nutritional status through measurement of a child's weight, age and height. World Health Organization (WHO) has developed new child growth standard (height-for-age, weight-for-age, weight-for-height and body mass index-for-age) in 2006 (5).

The underlying causes of under-nutrition are insufficient house hold food security, inadequate maternal and child care services, inadequate health services and unhealthy environment. Inadequate water supply and sanitation services are the major causes of child hood morbidity and mortality in Ethiopia (7).

The Government of the Federal Democratic Republic of Ethiopia developed National Health Policy in 1991. This policy has been the umbrella for the development of several initiatives and programmes include: National Nutritional programme (NNP), Health Sector Development Programme (HSDP) and Ethiopian Mini Demographic and Health Survey (EMDHS) to tackle child under-nutrition. Improving nutritional status is one of the strategic

objectives of HSDP-IV. The NNP developed and launched in 2008 for the aim of improving the nutritional status of children by implementing different strategies and initiatives especially at community level (8).

The 2014 EMDHS was conducted under the aegis of Federal Ministry of Health (FMoH) and implemented by Central statistics Agency (CSA). It collected data at facility level from the year 2010/11-2014 and provides the baseline to monitoring and evaluating the progress of child nutritional status that has been planned in HSDP-IV (10).

Generally, the aim of this project is building a model that helps to predict child under-nutrition by using EMDHS data. Data mining technology helps to identify hidden patterns or knowledge which helps for public decision making activities such as early controlling the occurrence of under-nutrition among under-five children.

1.2 Statement of the Problem

The major health problem of children in Ethiopia is largely preventable communicable disease and nutritional disorders (2). Under-nutrition contributes over half of (53%) child deaths and 10% loss on the Growth Domestic Product (GDP) in Ethiopia. Much progress has been made in addressing wasted category of under-nutrition through Expanded Outreach Strategy (EOS) for child survival, yet stunted category of under-nutrition has been neglected due to this it has consequences on children's physical and mental health problem(3).

Since, child under-nutrition is a major public health problem; it cannot be solved without understanding its causes, associated factors and consequences (4). The causes of under-nutrition may not similar in different geographical location and over a period of time. However, little information is available on nutritional status of demographic and socioeconomic segments of the country to formulate targeted tackling solution and overcome under-nutrition of under-five. Even though, the problems of the nutritional status of the children fairly documented, its specific determinants or associated factors have been hidden and remain poorly understood in Ethiopia.

Several researches done in some parts of Ethiopia show associated factors which contribute to the occurrence of child under-nutrition are mother's related factors, child's related and household's related factors. Even though, these studies have identified the factors contributing to under-nutrition, they have not analyzed the factors for the purpose of national and regional planning and intervention.

These studies have also shown that identified factors bring higher risk of anemia, diarrhea, fever, respiratory infections, under-nutrition and death on under-five children throughout the country. These problems in turn lead to under-five morbidity and mortality. Due to this high rate of morbidity and mortality, there will be an increase in the rate of school repetitions, school dropouts, and reduced physical capacity. Therefore, under-nutrition of under-five impacts the economy of families, health system and the country.

Despite the major strides to improve the nutritional status of children as planed in HSDP-IV; Ethiopian children still face the major problem of under-nutrition. This shows that FMoH couldn't achieve the objective of its own strategic plan. The 2013/14 HSDP-IV annual performance report of FMoH shows fewer declines of all kinds of child under-nutrition problem. Based on this report, the following points were identified as major challenges which includes; low attention on nutrition at different level, inequitable nutritional supplies - from national to community level, weak multisectoral nutrition coordination, inability to focus NNP implementation on the optimal factors of under-nutrition. These problems make ineffective planning that results poor achievement.

FMoH and CSA in collaboration with non-governmental organizations collected large a amount of data in 2014 through mini DHS. Due to the large amount of data generated on nutritional status; there is difficulty of deciding the attributes that highly contributes to the occurrence of under-nutrition particularly in children less than five years of age. So, there is a high need for application of data mining techniques in the big data of mini DHS.

Thus, to solve the sound problems related to under-nutrition of under-five children in Ethiopia, several activities have been carried out by the principal investigator through collecting , selecting and identifying attributes that highly contribute to the occurrence of under-nutrition, preparing nutritional status data for building the model, selecting and applied classification algorithms, building the model and evaluating discovered knowledge. At the end, this project designed a predictive model by using data mining tool as strategy which helps to make intervention for solving under-nutritional problems among under-five children in Ethiopia.

1.3 Objectives of the Project

General Objective

The main objective of this project is to build model for Predicting under-nutrition occurrence among under- five children in Ethiopia.

Specific Objectives

- To identify and describe factors those contribute to the occurrence of child under-nutrition at national level.
- To prepare data for building model by extracting, analyzing, cleaning, and transforming it into a format suitable for data mining algorithm.
- To build the model by using data mining classification algorithms on cleaned under-nutrition dataset of the 2014 EMDHS.
- To evaluate the performance of model based on classifiers' evaluating criteria in predicting under-nutrition of under-five children

1. 4. Significance of the Project

This project work provides effective models to predict the occurrence of child under-nutrition in different regions of Ethiopia. The newly built models have numerous benefits to FMoH, regions and stakeholders for intervention of child under-nutrition problems.
Obviously, Ethiopian Government has been striding to improve child nutritional status for a long period of time. In addition to this, the products of this project can be used by FMoH and regional health Bureaus for making better decision regarding early detection and prevention of under-nutrition.

FMoH can also use these models for evaluating of HSDP-IV report as well as supporting equitable nutritional logistic distribution throughout the country. These models support health planners to understand the nature and patterns of the occurrence of under-nutrition in different regions of Ethiopia. Additionally, the end product of the project can be used as a baseline that helps as a reference for conducting further research and project work in the future.

Generally, the products of this project add a clue for intervention of under –nutrition, and then improve quality life of individual as well as economic status of the community.

1.5. Scope and limitation of the project

1.5.1 Scope

The scope of this project is limited to build the model that can assist in predicting under-nutritional status of the under-five year age children in Ethiopia using the 2014 EMDHS data set and data mining tool. The inclusion criterion of this project is records of children whose age group is under-five age year. The constructed model is limited to evaluate the child nutritional status progress towards strategic objectives of health sector development programme IV.

1.5.2 Limitation of the project

The project had limitations which listed as the following:

- The project used small data set, making the performances of classifiers relatively less.
- This project applied Weka 3.7.10. This version of weka does not have SMOTE technique which helps to balance imbalance classes. Therefore, the prediction is biased to the majority classes rather than being proportional with other classes.

CHAPTER TWO

2. Literature review

An attempt was made in this chapter to discuss the concepts used in this project and review the available literatures on the subject of this work. This chapter highlights three kinds of concepts and ideas. These are child under-nutrition, data mining and related works.

2.1. Under-nutritional Status

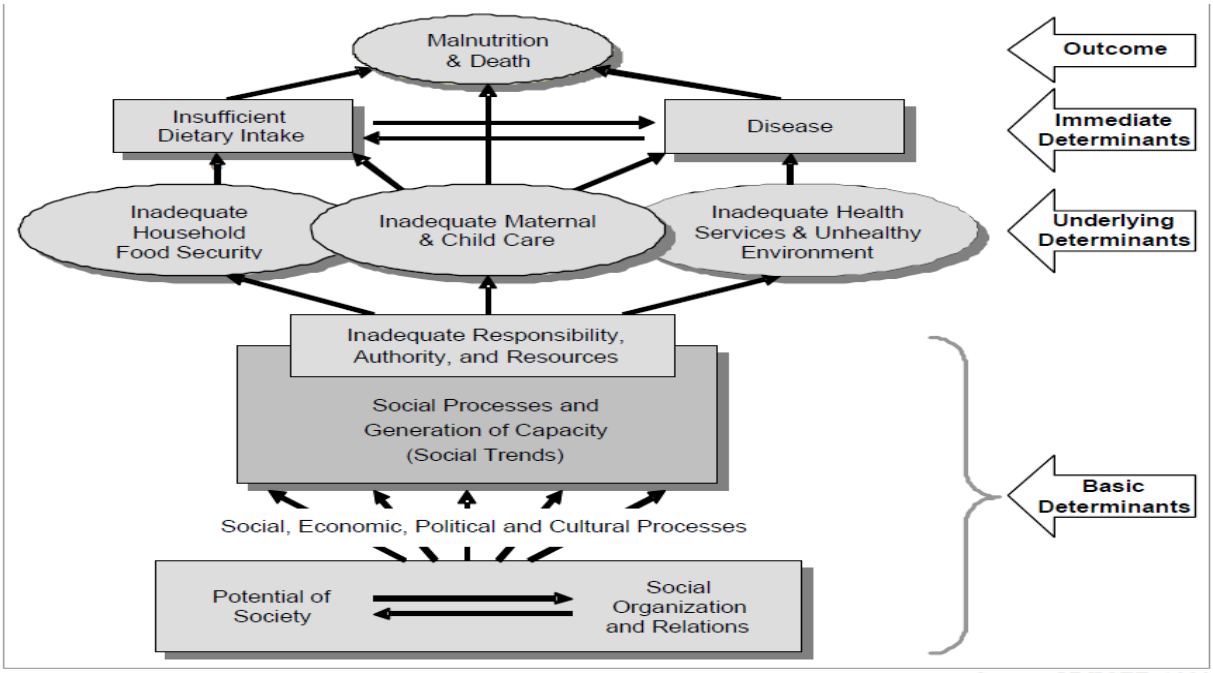
2.1.1. Overview of under-nutrition

Under-nutrition is the category of malnutrition that includes; stunted, wasted and under-weight. It occurs when children inadequately access to Food, health services and care. Under-nutrition weaken the immune system then putting the children at high risk of mortality and morbidity (2).The interaction between under-nutrition and infection creates a potentially lethal cycle of worsening illness and deteriorating nutritional status. Understanding the immediate and underlying causes of undernutrition in a given context is critical to delivering appropriate, effective and sustainable solutions and adequately meeting the needs of the most vulnerable groups like under-five children (3).

2.1.2. The cause of under-nutrition

The United Nations International Children's Fund (UNICEF) outlined the natures and causes of child under-nutrition in a conceptual framework. The three main causes are; underlying, basic and immediate (2, 11).The following figure: 1 shows the global conceptual frame work of the cause of under-nutrition.

Fig: 1 global conceptual frame work of the cause of under-nutrition



Source: UNICEF, 1990

Immediate causes: As we seen from the above framework, both inadequate dietary intake and infectious disease are immediate causes of under-nutrition. They are mutual reinforcing factors for child morbidity and mortality (2).

Underlying causes: Sometimes these causes leading to immediate causes. It includes inadequate household food security, maternal and child care, health services and unhealthy environment (9). Maternal under-nutrition leads to poor fetal growth and low birth weight of child. Therefore, improve maternal care is very important to improve child nutritional status.

Basic causes: These causes are leading to underlying causes of under-nutrition. It includes; social, political, economical, ideological strictures that lead to lack of potential resource such as; financial, human, and social and natural resources resulted in underlying causes that contributes to the under-nutrition (8).At the family/household level, the heavy workload resulting from many household responsibilities

that women shoulder-such as food processing and preparation, firewood and water collection, and care for the sick (6).

2.1.3. Measurement of child nutritional status

The aim of measurements of child nutritional status is to understand children's health. Nutritional status of children was collected and assessed through the 2014 EMDHS by CSA under the authority of FMoH. This assessment was carried out throughout the country specifically, conducted at facility and household level by measuring anthropometric data: height, weight and age of children (5, 10).

Anthropometric data collected in 2014 was used to estimate child nutritional status. These anthropometric data: height, weight and age of children are used to formulate nutritional status indices. These formulated indices are height-for-age, weight-for-height, and weight-for-age. According to WHO standards; these indices are changed into z-scores and they are mentioned as height-for-age score (HAZ), weight –for-height score (WHZ) and weight-for-age score (WAZ). To come upon common understanding throughout the world, the WHO expressed three indices as standard deviation (SD). The three indices were briefly described as follow:

Weight-for-height index: It is used to show the present time nutritional status of children and formulated from weight and height of children. It gives wasted, overweight and obesity status to children based on the data. If WHZ of children fall below minus 2SD from the median of child growth standards of the WHO standard developed in 2006, the children are identified as wasted. If WHZ of children fall above 2SD of the median of WHO standards, the children are accepted as overweight or obese.

Height-for-age index: Both height and age anthropometric measurements are used to calculate the HAZ. According to WHO; this index helps to produce data on chronic stunted category of under-nutrition. If HAZ of children age 0-59 months fall under minus two standard deviations from WHO standards, the children are recorded as stunted. Stunted under-nutrition known as chronic malnutrition. It indicates long failures of food intake.

Weight –for-age index: by using this index, anybody can easily indicate nutritional status of the children. It is formulated from the integration of HAZ and WHZ indices. Both stunted

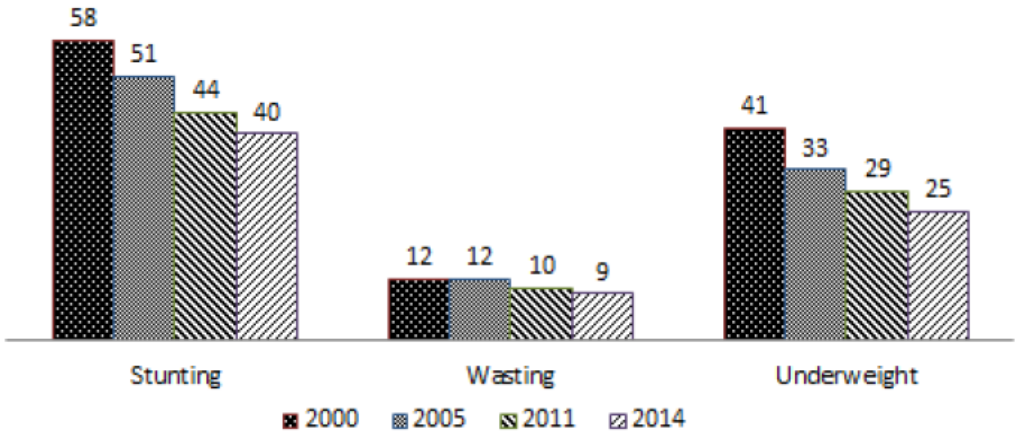
and wasted are determined by WAZ. It gives data on stunted and wasted under-nutrition. If WAZ of child whose age 0-59 months fall below minus two standard deviation from the WHO standards, the child can be considered as underweight. This child was accepted either stunted or wasted or both.

2.1.4. Trend's of Children nutrition status in Ethiopia (2000-2014)

The prevalence of under-nutrition of children under-5 is an indicator to measure progress towards Millennium Development Goal (MDG) , which aims to halve the proportion of people who suffer from hunger between 1990 and 2015 (2).

Figure 2 shows a downward trend in the proportion of children stunted and underweight over the four DHS surveys. The prevalence of stunting decreased by 31 percent (from 58 percent to 40 percent) between 2000 and 2014. The decline in the proportion of stunted Ethiopian children shows improvement in chronic malnutrition over the past fifteen years. The proportion of children underweight declined even more substantially by 39 percent over the same period. There was only a small decline in the prevalence of wasting over the last 15 years (10).

Fig: 2 trends in nutritional status of under-five from 2000-2014



Source: 2014Ethiopia Mini Demographic and Health Survey

2.2. Data mining

2.2.1. Overview of Data mining

Healthcare institutions are data intensive organizations that are generating and collecting large number of transactional data. Child nutritional status data is one of healthcare organization data(17).With the use of data mining techniques it is possible to extract useful knowledge and regularities that can be used in planning and decision making process in health care in order to improve work efficiency of child nutritional status.

Data mining is a new technology and uses a variety of data analysis tools. It can be defined as the process of discovering meaningful, new correlation, patterns, and trends by digging into (mining) large amounts of data stored in warehouse. It does not replace skilled business analysts or managers, but rather it gives them a powerful new tool to improve the job they are doing in Industries (22).

Data mining differs from traditional statistics in several ways: formal statistical inference is assumption driven in the sense that a hypothesis is formed and validated against the data. In contrast, Data mining is discovery driven in the sense that patterns and hypothesis are automatically extracted from data (19).

2.2.2. Data mining applications

> *Data mining is increasingly popular because of the substantial contribution it can make. It can be used to control costs as well as contribute to revenue increases. Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers (22).*

Health care has been the most rapidly growing segment of the nation's GDP for some time. Health facilities have collected large amount of data about patients, their health care problems, the clinical procedures used, their costs, and the outcomes. Understanding relationships in this data is critical for a wide variety of problems, ranging from determining what procedures and clinical protocols are most effective to how best to deliver health care to most people in an era of diminishing resources (19). Therefore, applying various data mining techniques for extracting and understanding such useful relationship in dataset is crucial in strategic planning and management of health care delivery.

2.2.3. Knowledge Discovery Process Models

Various Knowledge Discovery Process (KDP) models have been designed and implemented since 1990. KDP has a set of processing steps. At different time and places, researchers undertook knowledge discovery project by following knowledge discovery process model (KDP). Both industrial and academic models are the part of KDP model. Industrial model was developed after the deployment of academic model. According to Fayyad et al, they argued that researches are better carried out by following Academic models while projects can be executed by following the step of industrial models. Therefore, this project followed industrial model which is further described blow (20).

A. Industrial Models

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a part of knowledge discovery process model which was developed in 1990 for academic purpose of knowledge discovery project. The unique behaviors of this model are easy-to-understand vocabulary and the best for documentation. CRISP-DM is partitioned into sub-steps that provide all necessary details. The **CRISP-DM KDP** model has six steps. Each steps are briefly explained with relation to the project as follow (20, 21):

1. *Business understanding*. Objectives of the business were understood by the researcher and important requirements were identified based on business perspectives. Generally, at this stage all tasks were carried out through the following subsequent activities: determination of business objectives, assessment of the current situation, and determination of DM goals and generation of a project plan.
2. *Data understanding*. At this step of CRISP-DM, concerned data (both instances and attributes) were understood through carrying out several activities such as collection of initial data, description of data, exploration of data, and verification of data quality. The researcher performed any activities by moving forward and back ward direction between business understanding and data understanding.
3. *Data preparation*. At this step, data preparation tasks were started from data collection and ended by preparing final dataset. Data on nutritional status was collected from 2014 EMDHS at CSA by the researcher. The final data set which was loaded into DM tools was prepared at this stage. Several activities were carried out to build the final dataset. These

activities included data collection, data extraction, data selection, exploratory data analysis, data cleansing and transforming data into weka understandable format.

4. *Modeling*. At this point of CRISP-DM, various classification algorithms were selected and applied on the final data set for the purpose of building the predictive models. During performing experimentation for modeling, parameters of classification algorithms were calibrated until optimal values were obtained. The modeling tasks covers several activities including: selection of modeling techniques, generation of test design, creation of models, and evaluated of generated models.
5. *Evaluation*: After modeling activities were completed, the generated models were evaluated from a business objective perspective. The major activities carried out in this step include evaluation of the output, revision of the process, and deciding on the next step. The main objective of evaluation of generated models was deciding on an important business issues which have not been adequately observed. Finally, a decision about the use of DM outputs must was reached.
6. *Deployment*: This is the last step of CRISP-DM model. After modeling and evaluation were completed, the next step was either to deploy the new models or to understanding the business. At this step, the discovered knowledge should be summarized and put in a way it can be used by the concerned body. Based on the request, this step can be mainly about providing a report or implementing the extracted rules. If the implementation is requested, this is the step, where the extracted rules integrated into planning and used for evaluating the report. But this project providing a report to solve under-nutrition problems in future.

2.2.4. Data Mining Tasks

Data mining consists of many up-to-date techniques such as classification (decision trees, na¨ıve Bayes classifier, PART rule induction, k-nearest neighbor, and neural networks), clustering (k-means, hierarchical clustering, and density-based clustering), association (one-dimensional, multidimensional, multilevel association, constraint-based association). It can be divided into descriptive and predictive. While descriptive tasks have a goal on finding a human interpreted forms and associations, after reviewing the data and the whole construction of the model, prediction tasks tend to predict an outcome of interest (18).

Generally, both tasks have successful application requires data preprocessing, post processing (understandability, summary, and presentation), good understanding of problem domains and domain expertise.

2.2.5 Predictive Modeling

Predictive modeling is the task of building a concept model that expresses the target variable as a function of the predictor variables (18). The goal of predictive modeling is to minimize the difference between the predicted and actual values. A model representation consists of a set of *parameters* (attributes, operators and constants) arranged in some kind of *structure*. Predictive modeling is the process of *training* the parameters of the model using a data mining algorithm to fit a set of instances of the concept as well as possible. The instances that are used to build the model are consequently called the *training set*. A model can have a predefined static structure or it can be developed dynamically during training.

In predictive models, the values or classes we are predicting are called the *response*, *dependent* or *target variables.* The values used to make the prediction are called the *predictor* or *independent variables* (22). 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. In contrast, descriptive techniques such as clustering are sometimes referred to as *unsupervised learning* because there is no already-known result to guide the algorithms.

Classification

It aims 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. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern (22). Classification techniques of data mining consists of decision trees, na¨ıve Bayes classifier, k-nearest neighbor, PART rule induction and neural networks)(23)

A. Decision tree: It is a graphical representation of the relations that exist between the data in the database. It is used for data classification. The result is displayed as a tree, hence the name of this technique.

B. K-Nearest Neighbor (K-NN): K-Nearest Neighbor (K-NN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbor) and classified data points according to the voting system (23).

C. Bayesian Methods: This method is based on probabilistic knowledge. This method goes by the name Naïve Bayes, because it's based on Bayes's rule and "naively" assumes independence- it is only valid to multiply probabilities when the events are independent. Thus the naïve bayes rule outputs probabilities for the predicted class of each member of the set of test instance. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data. It is a simple classifier which is achieved by using classification algorithm (23, 24).

D. PART RULE Induction classification: Decision rule can 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 data set serve the same purpose as decision tree (27). Thus, a decision rules are also called as a classification rules (28), indicating that the rules can be used to predict the class of unseen instances.

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 consequent consists of the class value for the target attribute given by the particular leaf node (27).

2.2.6 Evaluation of predictive models

When a big dataset composed of known inputs corresponding to known outputs exists, generated model can be evaluated by splitting the available dataset into training and test set. Training set used for fitting the model, and test set used for evaluation of its goodness of prediction (20). How split this dataset into training and test set? There are three approaches

used to split such dataset. These are cross validation, holdout sample, and boots trap. They are the most popular in supervised learning used for accuracy estimation and model selection.

Cross-validation is better than bootstrap and holdout. In boots trap prediction error calculated on highly overlapping data such as training and test set. So that it has low variance, but extremely large bias on some problems. Computationally, bootstrap more expensive than cross-validation. However, in cross-validation prediction error was calculated on a portion data that was set for this purpose.

Cross-validation: It is an approach used for splitting a given dataset into training and test set for the purpose of estimating the predictive accuracy of classifier. It is used in situations where the data relatively small but difficult to split it into two parts (19, 20).

In stratified 10-fold cross-validation, the given dataset is divided into ten equal parts. Each part held out in turn and used to estimate the accuracy of the classifier, while learning algorithm is trained on 9/10 of the given dataset to fit the model. So that each fold is used once for testing and k-1 times for training iteratively.

Confusion metrics: After predictive model is developed, the model should be evaluated how it will perform for the future data that is not seen during the model building process (21, 30). It is a very useful tool for understanding results. Thus, to evaluate the performance of constructed predictive model, confusion matrix is very important tool.

This tool also called as contingency table. It shows how well the model predicts and presents the details needed to see exactly where things may have gone wrong. Also, it shows the counts of the true and predicted class values. As indicated in table: 2, the columns of the table indicate the true classes, and the rows indicate the predicted classes. Therefore the diagonal shows all the correct predictions.

Table;2 Simple confusion matrix

		PREDICTED CLASS		
ACTUAL CLASS		YES	NO	TOTAL
	Yes	TP	FN	TP+FN
	NO	FP	TN	FP+ TN
	TOTAL	TP+FP	FN+TN	TP+FN+ FP+ TN

Source: Data mining: Concepts and techniques. 3rd edition, 2012

Based on the above table, the performances of the classifiers were evaluated by calculating the following measurements. These measurements are accuracy, true positive rate (TPR), false positive rate (FPR), precision, recall and ROC curve.

Accuracy: The accuracy of the classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. It is estimated by dividing the total correctly classified positive and negative instances by the total numbers of samples.

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

Sensitivity: It referred to as true positive rate (TPR). It can measure how often what we find is what we are looking for .It is the proportion of positive tuples that are correctly identified.

$$\textbf{TPR=} \frac{TP}{TP+FN}$$

Specificity: It referred to as false positive rate (FPR). It can measure how often what we find is what we are not looking for. It is the proportion of negative tuples that are correctly identified.

$$\textbf{FPR} = \frac{TN}{TN+FP}$$

Precision: It measures exactness of the classifier. It indicates the percentage of positives which the learned model classified as actual positives. It can be estimated by dividing correctly classified instances by the total number of correctly and incorrectly classified instances. **Precision=** $\frac{TP}{TP+FP}$

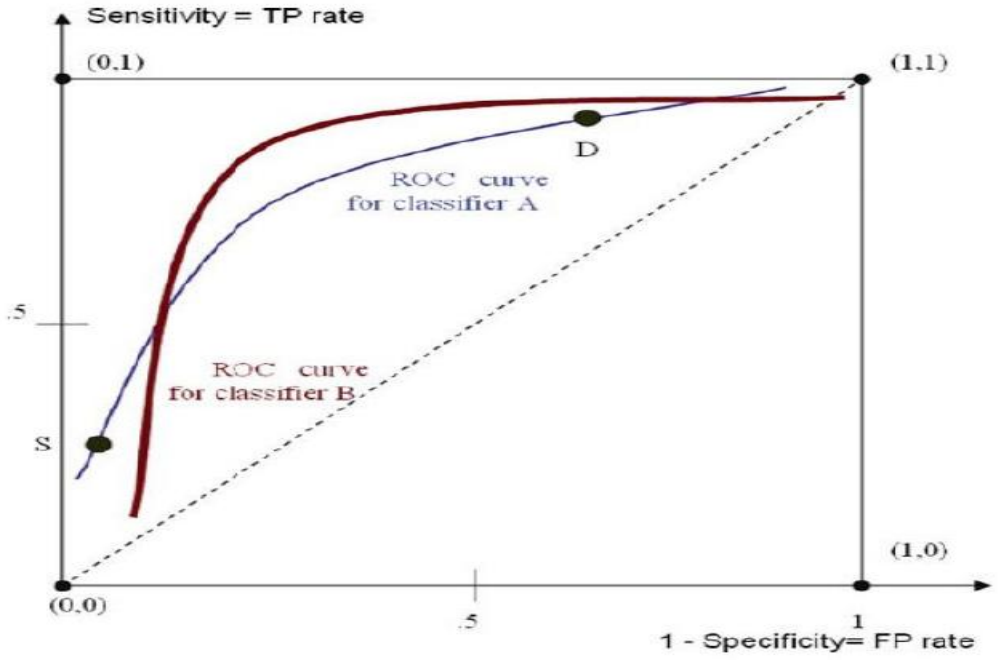
Recall: It measures completeness of the classifier. It shows the percentage of actual positives classified as positives by learned model. These measures can be computed as:

Recall= TP/TP+FN

Receiver Operating Characteristics (ROC)

This curve helps to compare performance of the classifiers and commonly used in medical decision making process. It is constructed by drawing curves in two dimensional spaces. These axes defined as the *TPR* and *FPR (18, 20)*. These axes are represented by using terms of sensitivity and specificity. As indicated in figure: 3, the y-axis represents Sensitivity = TPR, while the x-axis represents 1−Specificity = FPR. As illustrated in figure: 3, the values of TP rate plotted on the vertical axis while FP rate plotted on the horizontal axis. The performance of different classifiers with different parameters can be compared by inspecting the ROC curves. The larger area under the curve showed the better performance of classifier, while the smaller area under the curve indicated less performed classifier.

Fig: 3 Sample of ROC curve for classifiers



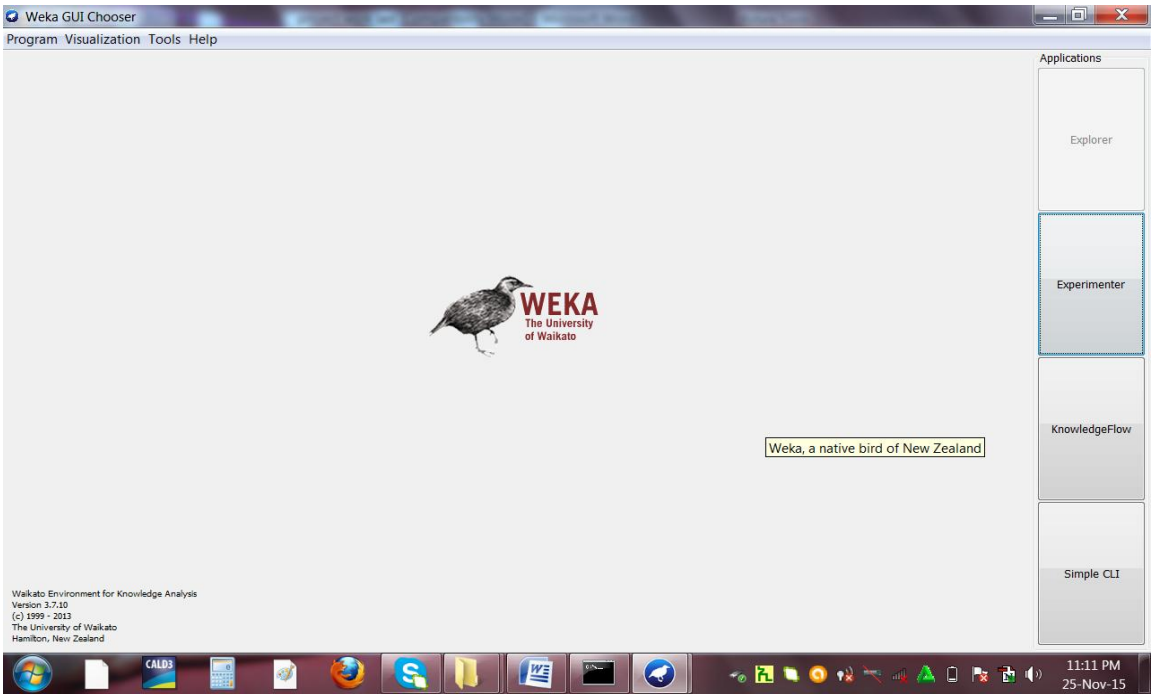
Source: Data mining: knowledge Discovery Approach, 2007.

2.2.7 Waikato Environment for Knowledge Analysis (WEKA)

Weka is a machine learning algorithms developed by department of computer science of Waikato University, New Zealand. Its name stands for Waikato Environment for Knowledge Analysis. It is open source software described in java and issued under the GNU General public license (29, 31).

Weka is a collection of several tools and run through user interface. These tools used for data pre-processing, classification, regression, association and visualization. There are two types of user interface in weka application such as GUI and SCLI. The three GUI found in weka are explorer, Experimenter and Knowledge flow. Since the aim of this project work is building the predictive model by using classification algorithm, all activities limited in GUI called explorer.

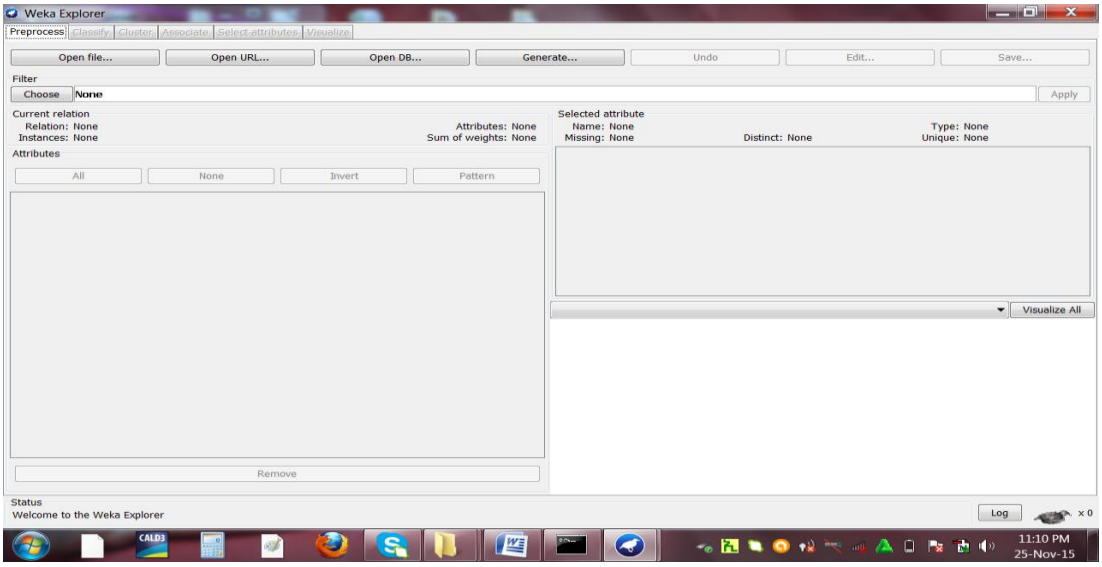
Fig: 4 Weka GUIchooser



Source: Weka version 3.7.10

Explorer: The easiest way to use weka is through graphical user interface called explorer (29, 31). This gives access to all of its facilities using menu selection. It guides the user by presenting choices as menus. As illustrated in Figure:5. At very top of explorer interface window, there are six tabs include: preprocessing, classify, cluster, association, select attribute and visualization. Of the six tabs, Classify tab is very crucial tab for prediction.

Figure:5 Weka 3.7.10 explorer window



Source: Weka version 3.7.10

Weka understandable file format: The data is often presented in a spreadsheet or database. However, both ARFF and CSV are common Weka's understandable file format (29). As mentioned in data preparation section of this project work, first data had obtained in SPSS format, then the data exported to Microsoft office excel for preprocessing. In order to make weka understandable file format, prepared data with excel file format was converted to CSV file format. Then, all tasks were performed by accessing data from its destiny.

2.3. Related work

Children are the most vulnerable group to under-nutrition in developing countries because of low dietary intake, infectious disease, lack of appropriate care, and inequitable distribution of food. Poor nutrition of children is continuous serious problem in Ethiopia. Some related researches have been supporting and explained briefly as follow:

1. ***Kimberly Moor.W*** (14) Conducted correlation hypothesis methods in sub-Saharan Africa in 2011, to examine the correlation between improving the quality of drinking water source and sanitation facilities and the likely hood that a child that will be stunted , wasted and under weight. Data variables of child and maternal characteristics, house hold specific variables of individual countries were collected from 2005-2011 by the demographic health and survey.
He used probit models methods for analyzing equations with binary variables. The probit models estimate predicted probabilities of a child being malnourished.

The first model predicts the likely hood of stunting and being under weight as a function of quality of drinking water and sanitation. The second predicts the likely hood of stunting, wasting and being under weight as a function of quality of water and sanitation and individual child-specific characteristics and mother-specific characteristics, including child age, whether or not the child is first born and mother's education. The third model adds house hold specific characteristics to the water and the sanitation variables including number of house hold members, number of children under five in house hold and house hold wealthy index. The fourth model combines models two and three to control for both child- and house hold specific characteristics.

The result of the study revealed that lack of adequate drinking water and sanitation facilities is correlated with child malnutrition and that increasing the quality of drinking water and sanitation facilities would correspond with a decrease with child malnutrition.

2. ***Tadiwos and Degnet*** (32) conducted cross-sectional study in 2013, in kombolcha District of Eastern Hararghe, Oromia regional state, in order to explore the key determinant of child malnutrition. They used two stage-sampling procedures to collect data from 249 under five years of age children. The collected data were analyzed and discussed using several descriptive statistics and logic regression model. The survey result revealed that, 45.8%, 28.9% and 11.2% of sampled children are stunted, underweight and wasted respectively. Also estimated result indicated that child nutritional status strongly associated with the child's age, gender, immunization status and the mother's use antenatal care, farm size, water source, latrine use and incidence of morbidity.

3. ***Mengistu*** (33) conducted cross-sectional study in 2012 to assess the prevalence of malnutrition and associated factors among children aged 6-59 months at Hidabu Abote District, North shoa, Oromia. Children were selected from each kebele by simple random sampling .Anthropometric measurement and structured questioners were used .Data were processed by using Epi-info soft ware and analyzed by SPSS soft ware.

His study revealed that, 47.6%, 30.9% and 16.7% of children were stunted, underweighted and wasted respectively in the study area. The main associated factors of stunting were found to be child age, family month income, and children were received butter as pre-lacteal feeding and family planning. Under weight was associated with number of children households and children were received butter as pre-lacteal feeding. Treatment of water, in households the only variable associated with wasting in the study area.

4. ***Asegedech***(34) conducted descriptive statistical analysis in 2014 by applying logit model method to analyze the determinants of child malnutrition among under five children of farming households in central zone of Tigray, Northern Ethiopia.

The result of her study revealed that the significant determinants of malnutrition are ,House hold age, Education of house hold, the presence of latrine facility in the house hold, use of treated water, sex of child, Child age and child birth intervals. The study showed that the prevalence of nutritional status among under-five year children are, 5.9%, 51.18% and

18.9% were wasted, stunted and underweight respectively. Consider to environmental aspect, 80% households access to safe water and 72% coverage of latrine facilities in the households.

5. ***Zeneba Merkos*** (35) conducted a research on application of data mining to predict nutritional status of under-five children using EDHS 2011 data set on 9,607records and 17 variables with the purpose of identifying the factors of affecting under-five malnutrition. The methodology employed to perform the research work was hybrid model. He used weka soft ware 3.6.8 version to extract the hidden patterns among the variables under the study.

To build predictive models he used J48, Naïve bayes and PART rule induction algorithms. The predictive model developed using PART pruned rule induction found to be best performing having 92.6% of the accurate result and 97.8% WROC area. In general, the results from this study were encouraging; it can be used as decision support aid for health practitioner.

The above several works are related to this project work. They were done in different geographical location and periods. The government of Ethiopia has prepared a successive strategic planning known as HSDP in every five years since 1997 GC. In such a way, DHS that serve as one of a data source for evaluating the achievement of HSDP targets has being continuously conducted in every five years by CSA. Since no any previous projects or researches conducted on Ethiopian 2014 mini DHS, this project was aimed to support the above purpose of DHS by discovering patterns using data mining techniques. This would help in planning a better strategy of interventions for under-five children's nutrition.

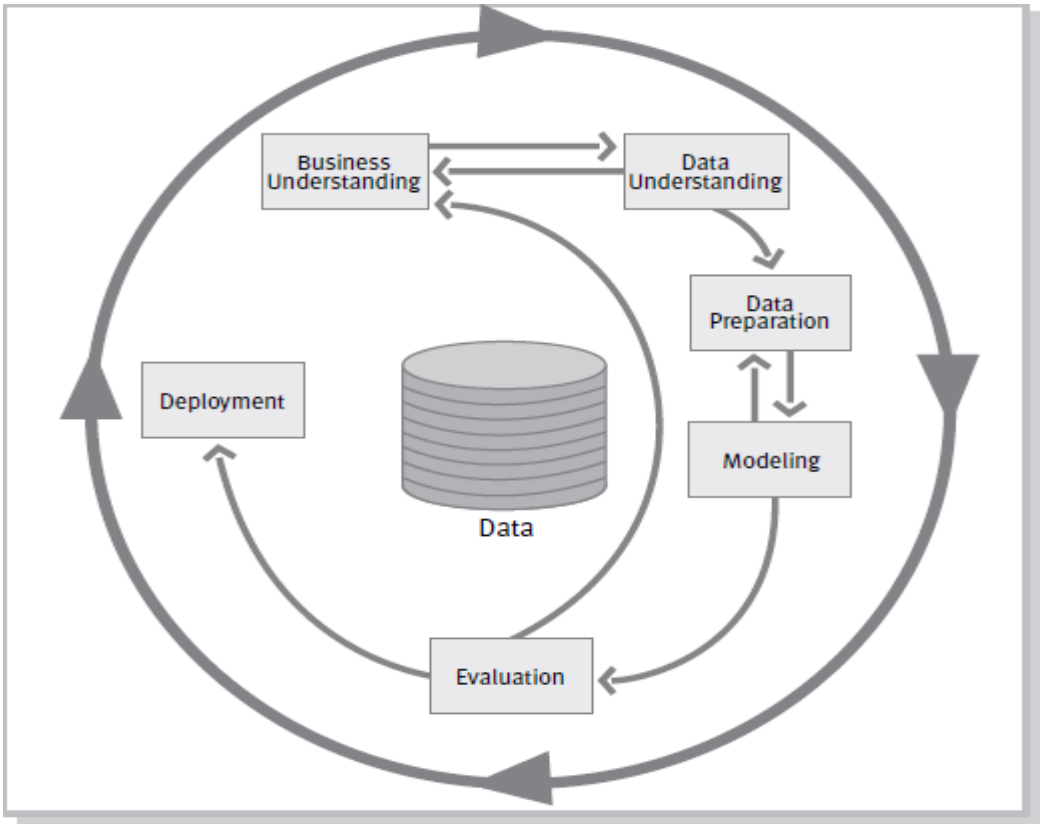
CHAPTER THREE

3. Method

3.1. Study design

To achieve the objective of this project, CRISP-DM Model was used since it is easy- to-understand vocabulary and good for documentation. It is one of the most widely used data mining methodology for knowledge discovery that consists of six steps; Business understanding, Data Understanding, Data preparation, modeling, evaluation and deployment of the model (20,21). The six step of this model is illustrated in figure 6.

Fig: 6 CRISP-DM models



Source: http://www.CRISP-DM.ORG

3.1.1. Business understanding:

In this step, assessing and understanding the business of FMoH aimed to determine its objectives /targets related to child nutrition from the perspective NNP. Discussion for understanding the problem of child under-nutrition was conducted among domain experts of maternal and child health-at FMoH. At the last of this step, understanding the business objectives and organizational review aligned with the data mining goal that is building the predictive model.

3.1.2. Data understanding.

The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of available data (20). The main aim of understanding the existing data is to create the target dataset with selected variables. The primary source of data for this project work was 2014EMDHS dataset of CSA. The database contains data on Child record, individual records, family health and maternal health.

For this project work, child records on nutritional status of under- five have been selected. Basically, child nutritional status dataset contains 5,579 instances and 440 attributes. From this dataset 5,177 instances and 19 attributes were selected with discussion of domain experts of FMoH. These attributes are: Region, residence, sex of child, age of child, birth interval, parity, mother's age, mother's educational level, age of house hold, sex of house hold head, number of ANC visits, place of delivery, source of drinking water, type of toilet facility, total member in house hold, household has radio, household has TV and wealth Index. These attributes were set as independent factors that are relevant to predict dependent variables such as stunted, wasted and underweight malnutrition using classification algorithms of DM. This data assist this project by providing up-dated key information of children, in order to predicting the occurrence of child malnutrition.

3.1.3: Data preparation.

This step of CRISP-DM covers all activities needed to construct the final data set (21). Data preparation tasks of this project work were started from initial raw data to prepared final data which will be fed into DM tools. The following activities were undertaken in this phase

include data extraction, collection, cleansing, exploratory analysis and data transformation. In order to correct the errors identified trough observation from the preprocessing stage-measures like filling the missing values based on the idea observing the neighbor records, correcting inconsistencies and removing instances with missing values were carried out. The cleaned data further processed with feature selection and extraction algorithms to reduce dimensionality by summarization of data.

3.1.4: Modeling

To build a model for predicting child nutritional status categories; WEKA 3.7.10 DM software was used. At this point the researcher prepared clean data in *ARFF* format for modeling purpose. Then, various modeling classification techniques were selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their parameters to optimal values. Based on literature review three algorithms of classification techniques were selected for modeling. These are J48 of decision tree, PART rule induction and Naïve Bayes. These algorithms were applied on the child nutritional status of 2014 EMDHS to build model for predicting child under- nutrition. At the final of this step, the best performed algorithm would be selected for building the model.

3.1.5: Evaluation

After one or more models have been built by classifier, the performance of each model was evaluated by using model evaluating criteria such as accuracy, True Positive Rate (TPR), False Positive Rate (FPR) and Receiving Operating Curve (ROC). Additionally, acceptability of generated rules was evaluated by understanding novel patterns, useful and interpretation of new discovered knowledge. 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 DM results should be reached. The key sub steps in this step include evaluation of the results, process review, and Determination of the next step.

3.1.6. Deployment/using discovered knowledge.

To organize and present the built model in a manner that owners and stakeholders can use it for planning and management of child nutritional service, the team was established and integrated the built model into routine decision making. Depending on the requirements, this step simply generating a report for further plan development, monitoring and management of child nutritional status from national to regional level.

3.2. Dissemination of Results

The report of this project would be presented and disseminated to various concerned bodies such as school of Information science and school of public health of Addis Ababa University, MoH and CSA by summiting original copies. The report will also be placed in the libraries of the University for those who are interested in the area to make further investigation and for reference purpose.

3.3. Definition of terms

Nutrition: The human being blood required with protein, fat, carbohydrate and micronutrients needed for individuals for sustainable healthy life (2).

Under nutrition: the outcome of insufficient food intake, inadequate care and infectious diseases. It includes being underweight for one's age, too short for one's age (stunting), dangerously thin for one's height (wasting) and deficiency vitamins and minerals (micronutrient deficiencies) (16).

Stunting: It reflects chronic under nutrition during the most critical periods of growth and development in early life. It is defined as the percentage of children aged 0 to 59 months whose height for age is below minus two standard deviations (moderate) and minus three standard deviations (severe stunting) from the median of the WHO Child Growth Standards(2).

Underweight: is a composite form of under nutrition that includes elements of stunting and wasting. It is defined as the percentage of children aged 0 to 59 months whose weight for age is below minus two standard deviations (moderate) and minus three standard deviations (severe underweight) from the median of the WHO Child Growth Standards (16).

Wasting: Defined as weight for height below minus two standard deviations from the median weight for height of the standard reference population. A child can be moderately wasted (between minus two and minus three standard deviations from the median weight for height) (2)

Severe acute malnutrition: It is defined as the percentage of children aged 6 to 59 months whose weight for height is below minus three standard deviations from the median of the WHO Child Growth Standards, or by a mid-upper-arm circumference less than 115 mm, with or without nutritional oedema (10).

Overweight: It is defined as the percentage of children aged 0 to 59 months whose weight for height is above two standard deviations (overweight) or above three standard deviations (obese) from the median of the WHO Child Growth Standards (2).

Data mining: It is the set of procedures and techniques for discovering and describing patterns and trends in data (18).

3.4. Ethical consideration

This project was conducted after getting permission from the ethical clearance committee of School of public Health of Addis Ababa University through the department of health informatics. To obtain required dataset, objectives of the study were explained to domain experts. Then required data set was obtained after getting permission from Ethiopia CSA.

CHAPTER FOUR

4. Business understanding and data preparation

4.1 Business understanding

Organizational understanding through discussing and interviewing with domain experts were carried out by the researcher for further understanding problems related to child nutritional status. Under-five children are most vulnerable of under-nutrition. Another group of concern is pregnant women, given that a malnourished mother is at high risk of giving LBW baby who will be prone to growth failure during infancy and early childhood, and be at increased risk of morbidity and early death (3).

The basic things of this project work were identifying determinants of under-nutrition. As discussion conducted with experts of FMoH determinant of child under-nutrition considered as predictors are: household's sanitation facility, source of drinking water, wealth index, age of household head, sex of house hold head, education level of house hold head, family size, sex of child, child birth interval, place of delivery, number of ANC visits and number of household members. Currently, all categories of nutritional status prevalence in the country are declining by 5%, but HSDP-IV objective of child nutrition is not achieved due to identified challenges in annual performance report of FMoH conducted in 2013/14 at national level. These challenges are Low attention on nutrition at different level, inequitable nutritional supplies from national to community level, poor quality of nutritional reports obtained from regions, which makes weak multisectoral nutrition coordination and regional linkages.

Reviewed documents depict that, the government strides to improve child nutrition problem. Several policies, programs, strategies and initiatives have been developed since 1991.These are Health policy in 991, HSDP in every five years, Data collected in every five years by CSA, Sanitation strategies in2006, NNP in 2011.

Major investments in child health in Ethiopia have yielded a substantial decline in under-five mortality rates; it is expected that the country will achieve MDG4. However, the last step is difficult unless the underline causes of child mortality are addressed. Under nutrition

is the main culprits causing of high child under-nutrition. It has an enormous impact on health, wellbeing and productivity. In both 2008 and 2012 the Copenhagen Consensus rated interventions to reduce under-nutrition of first priority among ten of the world's most important challenges. Therefore, addressing the problem of under-nutrition is a critical to achieving Millennium Development Goal (MDG) (6).

4.1.1 National objectives on child nutritional status

The government has already put in place program and initiatives at national level with set targets that directly and indirectly contribute to the reduction of under-nutrition. These programs include increasing agricultural productivity, promoting girls' education, Maternal and child care, Water, Sanitation and Hygiene (WASH) and NNP. The FMoH has being facilitated and supported the scale-up of these initiatives/programs in NNP to achieve the strategic objectives outlined below (7).

1. Reduce the prevalence of stunting from 44.4% (2011) to 30% by 2015
2. Reduce the prevalence of wasting from 9.7% (2011) to 3% by 2015
3. Reduce the prevalence of underweight from 28.7% (2011) to 21% by 2015

By discussing the above national strategic objectives, it aimed to aligning the built model with their achievement by supporting any action taken through NNP.

4.1.2 Determination of data mining goal

After determining objectives /targets related to child nutrition at the level of FMoH, data mining goal was set by the investigator as follow.

The main purpose of this project is to apply classification techniques of data mining tool on the 2014 EMDHS data to build model that predicts categories of child nutritional status among under-five in Ethiopia.

4.2. Data preparation process

The outcome of data mining and hidden knowledge discovery heavily depends on the quality and quantity of available data. Data understanding step focuses on creating a target data set 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 (20).

The original dataset of 2014 EMDHS is available in SPSS format. It is exported to excel file then converted in *ARFF* for WEKA 3.7.10 tool understandability.

As mentioned in methodology part of this project work, data preparation tasks include activities such data collection, data extraction, data exploratory analysis, data cleansing and data transformation were undertaken and explained as follow. The aim of this phase was preparing final data set which would be fed in data mining algorithms.

4.2.1 Data source

The source of data used for this project is obtained from 2014 EMDHS data set, which was conducted under the umbrella of FMoH and implemented by CSA in 2014. The survey provides health and Demographic indicators at the national and regional level for rural and urban areas. It is a national representative survey contains 5,579 records and 440 attributes or variables on children. It was collected from 2010/11-2014/15 at health facility. Data contains in EMDHS are; birth records, couple records, children records, house hold member records, individual records, HIV test records provided in SPSS file format. For this project child records were used.

4.2.2 Attribute and Instance selection

To decide on relevant attributes for building the model, the principal investigator has discussed with domain expert of FMoH. Also the researcher reviewed several literatures and collected related works which were done in different parts of Ethiopia. Out of four hundred forty (440) attributes in the data set, nineteen (19) attributes were selected for the purpose for building predictive model. These attributes are classified into two: Independent or predictor variable and dependent or predicted variable. Region, residence, sex of child, age of child, birth interval, parity, mother's age, mother's educational level, age of house hold, sex of house hold head, number of ANC visits, place of delivery, source of drinking water, type of toilet facility, total member in house hold, household has radio, household has TV and wealth Index, while, Nutritional status of child: wasted, stunted, under weight and normal are sub-class of dependent variables.

Irrelevances of instances were reduced and relevant instances were selected from the child records of 2014 EMDHS Dataset for predictive model construction. Out of 5,579 total child records, 5,177 instances were selected for classifying under-nutrition of children and 382 instances were excluded because, they were dead during data collection period. Also, 20 instances were reduced because they were obese children who were not project's target.

Table: 3, Description of selected attributes from 2014 EMDHS Dataset

Field name	Variable name	Data Type	Description	Values
B4	SEX	Nominal	Sex of child	Male, Female
HW1	AGE	Categorical	Age of child in month	Child age in month starting from 0-59 months
B11	BIRTH INTERVAL	Categorical	Child birth interval in month	Numeric starting from below 24 month
V013	MOT-AGE	Categorical	Mother's age in categorical	15-19, 20-24, 25-29, 30-34 35-39, 40-44, 45-49
V106	MOTHEDUC	Nominal	Mother's education Attainment	No education, Primary Secondary , Higher
M14	ANC VISITS	Nominal	Number of ANC Visits during pregnancy	No ANC visits 1, 2, 3, 4, 5, 6, 7, 8, 9, Don't known/ missing
M15	PLACE OF DELIVERY	Nominal	Place of delivery	10=Home, 11=respondent's home, 12=other home, 20=public sector, 21= Govn't Hosp, 22=Govn't HC 23=Govn't HS, 24= Govn't HP, 26=other public sector, 30= private sector, 31=Private Hospital, 32=Private Clinic 33= NGO's HF 36=Other private sector, 51=On the road, 96= other
V136	HOUSE HOLD MEMBERS	Numeric	Total of Household members	Numeric
V151	SEX OF HH HEAD	Nominal	Sex of household head	Male, Female
V152	AGE OF HH HEAD	Categorical	Age of household head	16 years and above
V190	WEALTHINDEX	Nominal	Wealth Index	poorest, poorer , Middle richer, richest
V113	SOURCE OF DRINKING WATER	Nominal	Source of drinking water	10=piped water ,11=Piped into dwelling, 12=piped into yard, 13=public tap, 20= Tube well water, 21=borehole, 30=Dug well, 31=protected well, 41=protected spring, 51=rain

				water, 71=bottle water, 32=Unprotected well, 42=Unprotected spring, 61=Tanker track,62= cart with small tank,40= surface water, 43=river/dam/lake/pond, 96=other
V116	TYPE OF TOILET FACILITY	Nominal	Type of toilet facility	10=flush toilet, 11= flush to piped sewer system, 12= flush to septic tank, 13= flush to pit latrine, 14= flush to somewhere else, 15= flush don't know where, 20= pit latrine toilet, 21=VIP, 22= pit latrine with slab, 23= pit latrine without slab, No facility, 31= No facility/bush/field, 41=composite toilet, 42= Bucket toilet, 43= hanging toilet, 96= other, 97= not de-jure
V024	REGION	Nominal	Region	Tigray, afar, Amhara, Oromia, somale, Benishangul, Gambela, Harari, SNNP, Addis Ababa, Dire- Dawa
V025	RESIDENCE	Nominal	Types of place of residence	Urban , Rural
HV2O7	HH- HAS RADIO	Nominal	Radio available	Yes, No
HV2O8	HH-HAS TV	Nominal	TV available	Yes, No

The above table shows independent variables that contribute the occurrences of dependent variables with their values .Independent variables those selected for purpose of predicting dependent classes such as stunted, wasted, underweight and normal.

4.2.3. Exploratory data analysis

Descriptive data summarization techniques used to classify the representative properties of data and highlight which data values should be treated as noisy or outliers. Description of selected attributes together with the exploratory data analysis has been performed by excel soft ware and SPSS application. After analysis had performed, cleaned and bad data were observed on frequency tables. As shown on tables 4, 5, 6 and 7 the exploratory data analysis was performed to detect bad data such as the attributes with missing values and wrong entries or noises and inconsistencies in values of attributes. So that based on this analysis, the researcher easily understand and correcting them by applying data preprocessing mechanism.

Region and residence: They are most important and selected as associated factors of child malnutrition. These factors indicate the location and residence of mothers and child. As shown in table 4 Region has nine possible values, while residence only two.

Table: 4 Descriptive statistics Summary of Region and residence of children and mothers

	Frequency	**Percent**
Region		
Tigray	459	8.2
Afar	633	11.3
Amhara	526	9.4
Oromia	685	12.3
Somali	665	11.9
Benishangul-Gumuz	474	8.5
SNNP	731	13.1
Gambela	421	7.5
Harari	363	6.5
Addis Ababa	217	3.9
Dire Dawa	405	7.3
Missing Values	o	O
Total	5185	100
Residence		
Urban	974	17.5
Rural	4605	82.5
Missing Values	0	0
Total	**5177**	**100**

The above table: 4 indicated that, most of the data (13.1%) is collected from the SNNP region and rural residence (82.5%) and the smallest data (3.9%) is collected from Addis Ababa city. There are no missing values observed in both.

Child related factors: It includes the attributes such as child age, sex and birth interval. The child age is the Numerical attribute from 0-59 months. The possible values of this attribute are classified into six category :< 6, 6-11, 12-23, 24-35, 36-47 and 48-59. Sex of child is nominal attribute with possible values of the sex male and female. The statistical distribution of this attribute is shown in the table 5

Table: 5 Descriptive statistics Summary of Child related factors

	Frequency	**Percent**
Sex of child		
Female	2502	48.2
Male	2683	51.7
Missing values	0	0
Total	5177	100
Age of child		
<6	179	3.5
6-8	115	2.2
9-11	140	2.7
12-17	595	11.5
18-23	414	8.0
24-35	1478	28.5
36-47	1153	22.2
48-59	1111	21.4
Missing values	0	0
Birth interval of child		
below two year	1140	22.0
two to four year	2282	44.0
above four year	1751	33.8
Missing values	12	0.23
Total	**5177**	**100.0**

The above table: 5 indicated that the distribution of the instances among the values of the children's age group of 48-59(23.4%) is large age categories while; 6-11(6.4%) is small age category. The distribution of instances among the value of the sex of the child is nearly equal: male (51.6%) and female (48.4). There were no missing values observed in both sex and age of child's attributes. But there were missing values observed in birth interval of child.

Mother related factors of child malnutrition: these factors including; mothers education level, number of ANC visits and place of delivery. Mother's education is associated factors of the occurrence of child malnutrition, which directly effect on child care practices. It is nominal attribute that contains four distinct values (No education, Primary, secondary, and higher).

Table: 6 Descriptive statistical summaries of mother related factors

	Frequency	**Percent**
Mother's age		
15-19	184	3.5
20-24	964	18.6
25-29	1679	32.4
30-34	1121	21.6
35-39	810	15.6
40-44	291	5.6
45-49	136	2.6
Missing values	0	0
Total	5177	100
Mother's education		
Higher	48	.9
No education	3686	71.09
Primary	1279	24.7
Secondary	153	3.0
Missing values	19	0.3
Total	**5185**	**100.0**
Number of ANC Visits		
No ANC visits	2883	55.6
1-3 visits	851	16.4
4 and above visits	1443	27.8
Missing values	9	0.17
Total	**5177**	**100.0**
Place of delivery		
Health facility	871	16.7
Home/Else where	4255	82.1
Missing values	51	1.0
Total	**5177**	**100.0**

As shown on the above table 6, place of delivery is high out of health facility and very less comparing at health facility. Mother age is largely dominated in the year interval 25-29 (31.2%). Number of ANC visits and place of delivery have missing values 0.17% and 1% respectively.

Household related factors: Age of household head associated to the occurrence of child malnutrition, older age of household age higher susceptible occurrence of child malnutrition. The household head age category grouped into sex: 16-25, 26-35, 36-45, 46-55, 56-65 and above 66. Wealth index is economic determinants of child malnutrition at household level. It helps to indicate the level of expenditure and income of households. The wealth index attribute has the possible values are Poor, middle and rich.

Table: 7 Descriptive statistical summary of Household related factors

	Frequency	**Percent**
Age of Household head		
16-25	448	8.6
26-35	2074	40.0
36-45	1776	31.8
46-55	545	10.5
56-65	300	5.8
66^{+}	160	3.1
Missing values	0	0
Total	**5177**	**100**
Sex of household head		
Female	657	12.7
Male	4520	87.3
Missing values	0	0
Total	**5177**	**100.0**
Total living children in HH(parity)		
1-2	1618	31.2
3-4	1669	32.2
5^{+}	1878	36.2
Missing values	20	0.3
Total	**5177**	**100.0**
Total member of in HH		
1-3	356	6.9
3-5	2148	41.4
6-8	2043	39.4
9-11	490	9.5
12^{+}	142	2.7
Missing values	0	0
Total	**5177**	**100.0**

Household wealth index		
Middle	826	15.9
Poor	3004	57.9
Rich	1347	26.1
Total	5177	100.0
Source of Drinking water		
Improved	2341	45.1
Unimproved	2712	52.3
Other	70	1.4
Missing values	54	1.4
Total	5177	100.0
Type of toilet facility		
Improved	132	2.5
Non-improve	4956	95.6
Other	63	1.2
Missing values	26	0.5
Total	**5177**	**100.0**

The above table depicts that household age most dominantly occurred in the age interval 26-35 (40%), few of them lies above 66 age category (3.2%).There are no missing values. The most of mothers (52.6%) are poor, but (31.7%) of them are rich. The majority of the residences obtain drinking water from public tape/standpipe (25.5%), but a few of them obtain from tube well/borehole (0.3%) and there is no missing value. Most of the households use toilet services in bush/field (46.2%) because they have no toilet facility and the least number of household flush to somewhere else and its missing values only one.

4.2.4 Data cleaning

The dataset obtained from the CSA, on which the project conducted on have missing values and outliers (35). Attributes those have missing values are birth interval, place of delivery, education level of mother, parity, ANC, source of drinking water and type of toilet facility. Therefore, data cleaning activities have been applied in order to clean data by filling missing values, correcting noisy and resolving inconsistencies.

Handling missing values

Missing values is one or more fields of attributes which have no value in it. The existence of many such cases makes datasets incomplete and building models of any type and makes the model none representative of reality (35). Missing values and its problem are very common in the data cleaning process. Many data sets have a big problem of missing values. The missing values may me happen due to various reasons such as incomplete data entry, incorrect measurements, equipment errors, and lack of consistency with other recorded data.

To handle the missing values different techniques described as follow:

1. Replacing missing values by considering the nearest neighbor.
2. Replacing the missing values by most frequent values randomly

Table: 8 Summary of descriptive statistics of missing values

No	Attributes	Valid	Missing values in %	Type of bad data	Mechanism of handling
2	Birth interval	5173(99.7%)	12 (0.23%)	Missed	Replacing
3	Place of delivery	5134(99%)	51 (1%)	Missed	Replacing
4	Mother's education	5166 (99.6)	19 (0.3)	Missed	Replacing
5	Parity	5165(99.7%)	20% (0.3%)	Missed	Replacing
6	Source of drinking water	5131(98.9%)	54 (1.04%)	Missed	Replacing
7	Type of toilet facility	5159(99.5%)	26(0.5%)	Missed	Replacing

The above table: 8 depicts that summary of missing values of the attributes ranging from 12(0.23%) to 54(1.04%).

4.2.5 Data Transformation

It is about transforming the data to make it appropriate for mining. Hence, the researcher has encoded data variable based on the dataset as shown on the table 9.

Table: 9 Summary of data codification

Attribute name	*Old value*	*Codify*	*New value*
Age of child	Age between 0-59 month	Categorized into eight based on WHO child age categorizing	<6 , 6-8 , 9-11, 12-17 **,** 18-23 24-35 36-47, 48-59
Birth interval	In numeric	Categorized into three ;Below 24month, 24-47 month, above 48month	<24month, 24-47month, >48month
Place of delivery	10=Home, 11=respondent's home, 12=other home, 20=public sector, 21= Govn't Hosp, 22=Govn't HC 23=Govn't HS, 24= Govn't HP, 26=other public sector, 30= private sector, 31=Private Hospital, 32=Private Clinic 33= NGO's HF 36=Other private sector, 51=On the road, 96= other	(20, 21, 22, 23, 24,26,30, 31, 32, 33, 36) = **Health facility** (10, 11, 12) = **home** (51)**= on road** (96)**= other places**	Health facility Home/else where
Type of toilet Facility	10=flush toilet, 11= flush to piped sewer system, 12= flush to septic tank, 13= flush to pit latrine, 14= flush to somewhere else, 15= flush don't know where, 20= pit latrine toilet, 21=VIP, 22= pit latrine with slab, 23= pit latrine without slab, No facility, 31= No facility/bush/field, 41=composite toilet, 42= Bucket toilet, 43= hanging toilet, 96= other, 97= not de-jure	(10, 11,12,13,21,41,) = Improved (14, 15,20, 22, 23,42, 43, 96, 97)=non improved (31)=field/bush/open	Improved facility unimproved facility field/bush/open
Age of household head	16 years and above	In numeric starting from 18 years old	<18yrs, 19-24, 25-30, 31-36, 37-42, 43-48, 49-54, 55-60, 61-65, 65+
Parity	Numeric	In numeric starting from 0	0=0, 1-2=1, 3-4=2, 5and above=3
Members in house hold	Numeric	In numeric starting from 0	1-3=1, 4-6=2, 7-9=3, above 10=4
ANC Visits	None, 1-3, 4and above, don't know	(None)=0, (1-3)= 1, (4and above)=2, (don't know)=3	No ANC visits, 1-3 , 4& above
Nutritional status of under-five			
HAZ	Numeric	-<2SD	Stunted
		-2SD-2SD	Normal
		>2SD	Over
WHZ	Numeric	-<2SD	Wasted
		-2SD-2SD	Normal
		>2SD	Obese
WAZ	Numeric	-<2SD	Under weight
		-2SD-2SD	Normal
		>2SD	Overweight

4.2.6 Description of Preprocessed data

Generally, this original data obtained in SPSS format and then exported to Microsoft excel for preprocessed. Different corrective measures were undertaken on the original noisy data in order to handling missing values. As shown on table 10, the final dataset ready to be used for the data mining purpose contains nineteen variables and five thousand one hundred seventy seven instances. After all preprocessing had finished, cleaned data would be saved as CSV and automatically converted into *ARFF* format for WEKA understandable. These selected data set shown in table 10.

Table: 10 Summary of selected data set

Parameters	Original data set	Target data set
Total number of records	5,579	5,177
Total number of attributes	440	19
File format	SPSS16.0	ARFF

As we seen on the table 10 the description of dataset is ready to be imported to the data mining tool WEKA 3.7.10. After importing the CSV file to WEKA, the experiments explained in the next chapter.

CHAPTER FIVE

5.1 Experimentation and Evaluation of the model

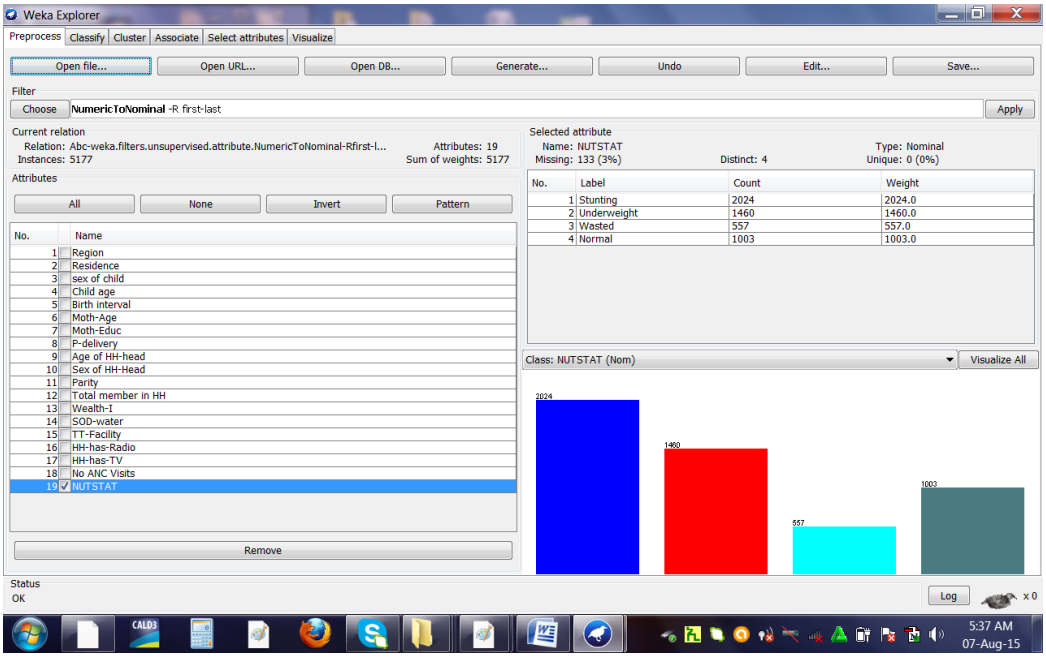
After finishing data preparation, the investigator made experimentation to build a model that helps to predict the occurrence of child under -nutrition among under-five year age children. The experiments were carried out with the use of classification algorithms such as Decision (tree J48), Naïve Bayes and PART rule induction. They are available in WEKA 3.7.10. To build the model by experimentation, Cross-validation was used and the above algorithms were applied on preprocessed dataset. This dataset contains 5177 records and 19 attributes used as training and test. The performances of each algorithm were evaluated with its own measuring criteria.

5.1.1 Experimental setup

In this section of experimentation, WEKA 3.7.10 and preprocessed child nutritional status dataset mainly used for experimental setup. As mentioned in data preparation section of this project work, the preprocessed dataset was saved in CSV file format in order to understandable by WEKA. Then, this dataset which saved using CSV format was imported into WEKA 3.7.10 by clicking on "open" button of explorer window and simply browsing it easily from file destination. After data is uploaded in WEKA, *Automatic operations by filters like numeric to nominal* were applied on all attributes; then they have the same nominal values.

The algorithms used for experimentation were found in WEKA 3.7.10.All experiments that were performed and discussed in the following subsequent sections were carried on 5177 instances and 19 attributes. These attributes are classified into independent and dependent attributes. As shown in fig 7, the dependent attribute or class attribute used in discovery process is nutritional status and the rest are independent attributes.

Fig: 7, WEKA 3.7.10 explorer window showing the list of attributes



As shown in fig: 7 independent attributes (predictors) were showed on the left side of the explorer window, while dependent (predicted class) were showed by bar graph at right side of explorer window. Dependent class categories counted and indicated; 2024 children stunted, 1460 children underweight, 557 children wasted and 1003 children were normal.

5.1.2 Attribute selection

It is the method used to search the relevant attributes in the given dataset which are the best work for building the desired model. There are two ways of searching attributes; manual based and automatic methods (28). Automatic method of attribute selection involves machine learning algorithms. Most machine learning algorithms are designed to learn which attributes are the most appropriate used for making decision.

Manual based method is the best way of selecting relevant attributes because it is based on deep understanding of the problem and what the attributes actually mean. As mentioned in business understanding phase of this project work, all attributes were selected by conducting very depth discussion with domain expert of the FMoH. Therefore, all attributes are the most significant for predicting under-nutritional status.

5.1.3 Classifier evaluation

In this project work classification algorithms were applied on a large dataset used for building predictive model. This dataset composed of 5177 instances and 19 attributes .As mentioned in literature review part of this project work, a big dataset is divided to training and test set for accuracy estimation and model selection.

Thus, dataset partitioning approaches such as cross-validation, holdout, and boots trap were discussed. Boots trap is more expensive in terms of computationally. It has low variance, but it has extremely large bias. Hold out is common to designate 2/3 for training and 1/3 for test set. In hold out approach more instances we leave for the test set, the higher the bias of our estimator. Based on the above justification, Cross-validation is better partitioning approach for this work.

Classification algorithms such as Naïve Bayes, J48 and PART RULE induction were applied on the same preprocessed dataset. During each experimentation, stratified 10-fold cross-validation splits preprocessed dataset (5177) into training set (9/10=4659) and test set (1/10=518). Training set is analyzed by the above algorithms and learner model is built, while test set used for evaluating the performance of the model.

5.2 Experimentation to model child under nutrition status.

In this section of this project work, several experiments were performed to select the best performed classifier and then build the model by it. All performed experimentation in the following subsequent section is divided into two phases. Phase-one experimentations have been done by each algorithm (J48, PART rule and Naïve Bayes) in order to find best classifier. In phase-two the experimentation has been done by best classifier to build predictive model.

5.2.1 Experimentations with Naïve Bayes

In this section of experimentation, Naïve Bayes classifier used to build predictive model when applied on preprocessed data of child nutritional status of EMDHS. It followed 10-fold cross-validation approach in order to partitioning available dataset into training (4659)

and test set (518).This classifier has its own parameters. The investigator of this project tried to iteratively changed classifier's parameters and then experimented until the best model obtained.

This method based on assumption of probability and assumes all attributes are independent. The probability of co-occurrence of an attribute value together with a particular outcome value is computed. Then, the class of new instance will be computed by multiplying the probabilities of values the instance has assumed under each attribute. Table 11 depicts those experimentations by Naïve Bayes algorithm with all attributes.

Table: 11 Experimentations done by Naïve Bayes algorithm by modifying its parameters.

No.Exper	Schemes	Accuracy	WTPR	WFPR	Precision	WROC
1	Naïve Bayes	51%	51%	24.7%	46.3%	74.4%
2	Naïve Bayes-0	50.83%	50.8%	24.8%	46%	74.4%
3	Naïve Bayes-D	50.83%	50.8%	24.8%	46%	74.4%

Table: 11 depicts that the performance measurement for Naïve Bayes with default values of its parameters. In experiment 1, the value of displayModelInOldFormat by default "false" in this value of accuracy 51%, WTPR 51%, WFPR 24.7%, Precision 46.7%and WROC 74.4%. During the value of "displayModelInOldFormat" **false** changed to "**true"** on the models performance in experiment-2, the value of measuring criteria slightly reduced. In experiments 2 and 3 Naïve Bayes showed the same result in all measurements. Of the three experiments, experiment-1 showed the best performance. So that it is selected as the best performed.

5.2.2 Experimentations by PART rule induction

In order to build the predictive model of under-nutrition of under-five children, this classifier was applied on preprocessed child nutritional status data (5177) of EMDHS data. Partitioning of dataset into training and test set was carried out by using stratified 10 fold cross-validation approach.

This algorithm builds partial decision trees and reads a path from the root of the tree to the leaf to read a rule. The rule is AND end together to give a complete set of rules. PART has almost similar set of parameters with J48 algorithm that can be adjusted to build better model from data set.

This algorithm has its own parameters. These parameters include binary splits, confidence factor, debug, minNumObj, numFolds, reducedErrorpruning, seed, unpruned and useMDLcorrection. During experimentation these parameters were adjusted by the investigator to find the best classifier.

Table: 12 summary of experimentation with PART algorithm

Exp #	Schemes	Accuracy	WTPR	WFPR	Precision	WROC
1	PART-M2-C-0.25-Q1	57.8%	57.8%	18.5%	55.5%	79%
2	PART-M2-C-0.1-Q1	59.6%	59.6%	18.7%	56.5%	80.9%
3	PART-M2-C-0.05-Q1	59%	59%	19.3%	55.4%	80.9%
4	PART-U-M2-C-0.25-Q1	50%	50%	18.9%	52%	69.6%
5	PART-U-M2-C-0.1-Q1	50%	50%	18.9%	52%	69.6%

As indicated on the above table, five experiments were performed on the same 19 attributes and 5177 records. As confidence factors are decreasing with Binarysplit by default false; but the value of accuracy, TPR and WROC are increasing until confidence factor 0.05, in turn these values are decreasing after 0.05 confidence factor. After experimentation and evaluation completed, experiment#2; scheme PART-M2-C-0.1-Q1 is selected for its best performance.

5.2.3 Experimentation by J48 decision tree

J48 classifier is used for data classification. It was applied on preprocessed data of child nutritional status of EMDHS data. This data was divided into training and test set by using cross-validation approach. J48 is algorithm can work on multiple valued attributes. As it

was mentioned in data description section, the independent attributes that affect the occurrence of under-nutrition of under-five are multi-valued.

As we have seen from table 13, four experiments made to find better classifier by using default parameter setting and varying algorithms important parameters. These parameters are binary splits, confidence factors, minNumObj, subtree Raising and unpruned.

Table: 13 Experimentation with J48 by modifying its parameters

Exp #	Schemes	Accuracy	WTPR	WFPR	Precision	WROC
1	J48-C 0.25-M2	62.9%	62.9%	19%	58.6%	80.2%
2	J48-C 0.1-M2	63.5%	63.6%	18.3%	62.3%	82.2%
3	J48-C 0.05-M2	63.3%	63.3%	19.4%	60.3%	82.3%
4	J48-U-M2	53.7%	53.7%	18.5%	54.4%	73%

As shown in table 13, several experimentations were performed to develop model with higher performance measures by adjusting J48 parameters.

The first experiment was performed by taking the default values of all parameters. The default value for binary splits is 'false', collapse tree 'true', a confidence factor is '0.25', debug 'false' minNumObj is '2', subtree Raising is 'true' and unpruned is 'false'. This experimentation has accuracy62.9%, WTPR 62.9% and WROC 80.2%

In experimentation 2 and 3, the default values for parameters are the same with experiment 1, but the confidence factor is decreasing from 0.01-0.05. As it shown on the above table, the measurement of output such as accuracy, WTPR, Precision and WROC were slightly decreased from experiment 2 to 3.

In experiment # 4, unpruned parameter changed to "true" on confidence factor is 0.25. From all the above experimentation, Exp #2 has high accuracy, WTPR and ROC that are 63.5%, 63.6% and 82.2% respectively. Therefore, Exp #2 selected as the best performance of the classifier for building the model.

5.2.4 Selecting best schemes of different algorithm for child nutritional status modeling

After experiments performed by each classification algorithm on the above subsequent section of experimentations, then the best performed algorithms were selected for phase to experimentation. Thus, the objective of this project is identifying which DM algorithm performances best in predicting the occurrence of under-nutrition. Therefore, the experiments in the study were carried out with J48 decision tree, PART rule induction and Naïve Bayes. The dataset used in all experimentation are the same.

For accuracy estimation and model selection stratified 10-fold cross-validation approach was used. Model comparison was performed using performance evaluation matrix like true positive rate, false positive rate, precision, ROC area and accuracy of the model. The detail result of the best model selected from each algorithm of classification category is shown in the table 14

Table: 14 summary of the performance of the best model created by classification algorithms

Exp #	Algorithm	Schemes	Accuracy	WTPR	WFPR	Precision	WROC
1	**Naïve Bayes**	Naïve Bayes	51%	51%	24.7%	46.3%	74.4%
2	**PART**	PART-M2-C-0.1-Q1	59.6%	59.6%	18.7%	56.5%	80.9%
3	**J48**	J48-C 0.1-M2	63.5%	63.6%	13.3%	62.3%	82.2%

The above table depicts that J48 Classifier has achieved the best performance as compared to PART and Naïve Bayes algorithms. J48 with all attributes provided the best accuracy of 63.5% and WROC 82.2%. But, WROC measurement is the most preferable than accuracy in public health research. Therefore, model from J48-C 0.1-M2 is selected as the best model and the investigator used this model to develop/extract some relevant rules.

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.1 -M 2

Relation: children's nutritional status data of 2015EMDHS-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 5177

Attributes: 19

Region, Residence, sex of child, Child age, Birth interval, Moth-Age, Moth-Educ, P-delivery, Age of HH-head, Sex of HH-Head, Parity, Total member in HH, Wealth-I, SOD-water, TT-Facility, HH-has-Radio, HH-has-TV, No ANC Visits, NUTSTAT

Test mode: 10-fold cross-validation

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

J48 pruned tree

=== Stratified cross-validation ===

=== Summary ===

Number of Leaves: 173

Size of the tree: 214

Time taken to build model: 0.67 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	3290	63.5503 %
Incorrectly Classified Instances	1887	36.4497 %
Kappa statistic	0.4572	
Mean absolute error	0.2428	
Root mean squared error	0.3561	
Relative absolute error	68.5905 %	
Root relative squared error	84.6391 %	
Coverage of cases (0.95 level)	97.3537 %	
Mean rel. region size (0.95 level)	68.8768 %	
Total Number of Instances	5177	

=== Detailed Accuracy By Class ===

	TP R	FP R	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.018	0.002	0.500	0.018	0.035	0.079	0.732	0.224	Wasted
	0.656	0.178	0.591	0.656	0.622	0.464	0.812	0.597	Underweight
	0.872	0.301	0.651	0.872	0.745	0.558	0.831	0.690	Stunting
	0.491	0.066	0.676	0.491	0.569	0.480	0.864	0.657	Normal
Av/w	0.636	0.183	0.623	0.636	0.595	0.463	0.822	0.606	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10	102	425	20	a = Wasted
5	958	361	137	b = Underweight
4	145	1765	110	c = Stunting
1	415	162	557	d = Normal

5.2.5 Confusion metrics for J48 classifier

A confusion metrics is known as contingency table which size 4x4. In this project work, it is a useful tool for calculating how well predictive classifier can recognize sub-classes of under-nutrition. The confusion matrix for J48 shown in table: 15 illustrates that out of the total 5,177 records 10 records were correctly classified as "Wasted", 958 records were correctly classified as "Underweight ", 1765 records are correctly classified as "Stunted" and 557 records were correctly classified as" Normal".

Table; 15, Confusion metrics for J48 classifier

Wasted	Underweight	Stunted	Normal
10	99	449	8
3	949	374	134
4	136	1816	68
0	416	146	441

As shown on the above table, all records classified into correct and incorrect. From5177 records: 3207 (63.5%) are correctly classified, while 1887 (36.4%) are incorrectly classified.

5.3 Generated rules

In these experimentations; **J48** classifier has achieved relatively the highest in most of the algorithms' performance evaluation criteria (convectional machine learning metric) compared to PART rule and Naïve Bayes algorithm. Therefore, the model generated by J48 classifier with all 19 attributes was selected as the model that can predict the child nutritional status. As can be seen from Annex-B, J48 classifier generated rules which have 173 leaves and 214 size of the tree. According to discussion made with a domain expert of child health specifically; child health 11 rules were selected.

1. If No of ANC Visits = 1-3 visits AND Region=Gambela AND Child age = 36-47m AND Type of Toilet Facility=No facility/bush/field AND SOD-water =Improved AND sex of child = Female: **Stunted** (6.0/3.0)

2. If No of ANC Visits = 1-3 visits AND Region =Tigray AND Child age = 36-47m AND birth interval=<48m AND mother age=30-34 years : **stunted** (8.0/2.0)

3. If Child age = 36-47m AND Place of delivery = Home AND Wealth-I = Poorest AND Mother Education = No AND Total member in HH = 4-6 AND Sex of house hold head = Male AND Age of house hold head = 37-42yrAND Region = **Amhara**: then Stunted 6.0/2.0
4. If No of ANC visits = 1-3 visits AND Region= Dire Dawa then : underweight 184.0/81.0
5. If No of ANC visits=1-3 visits AND Region =Harari AND mother education= no education then underweight (38.0/20) or 66%.

6. If No of ANC visits=4 visits AND mother education= no education AND Region =Oromia AND child age=48-59m AND Residence=rural: underweight 58.0/11.0.
7. If No of ANC visits=4 visits AND Region =Oromia AND childe age=48-59 AND residence =urban: normal 6.0/2.0.

8. If No of ANC Visits= 4 visits AND region =Addis Ababa then Normal (35.0/12.0) or 74%
9. If No of ANC visits=no visits region =Addis Ababa and mother education= no education AND SOD water=improved then underweight 5.0/1.0
10. If No of ANC visits = no visits AND Region =Harari then wasted (10.0) or 100%
11. If No of ANC visits = no visits AND Region= Addis Ababa AND mother education= no education =wasted (2.0/1.0).

Decision tree of J48 is graphical presentation of the relation between data set exists in data base. Several rules were extracted by this classifier. These rules were presented in IF-THEN for the aim of readable and understandable any users. The numbers in brackets at the end of each leaf showed correctly classified instances while the numbers mentioned after the slash (/) showed wrongly classified. The capability of the models for prediction was shown by success ratio. Success ratio will be calculated from number before slash divided by number before slash plus number after slash. Some generated rules by J48 model with all relevant variables listed above and explained as follow:

As we observed from rules #1, 2 and 3: J 48 algorithm classified attributes as predictors of stunted under nutrition in Gambella, Tigray and Amhara regions.

In rule#1 attributes were classified as predictors of stunted under-nutrition are mothers' who visited health institution for one to three times, children whose age 36-47, households who hasn't toilet and female children are more likely induces stunted under-nutrition in Gambella region. This rule implies that 6.0 instances were correctly classified while 3.0 instances wrongly classified out of 9.0 instances. This rule revealed; stunted category of under –nutrition occur with 67% of success ratio in Gambella region.

In rule #2 attributes were classified as predictors of under-nutrition in Tigray region are mothers' who visited health institution for one to three times, children whose age 36-47, birth interval of children below 48 months and adult mothers are predictors of stunted under-nutrition in Tigray region. This rule implies that 8.0 instances correctly classified while 2.0 instances incorrectly classified out of 10.0 instances. This indicated that stunted under-nutrition occur with 80% of success ratio.

In rule #3: in Amhara region stunted under-nutrition was predicted from attributes such as children whose age 36-47 months, mothers delivered at their home, uneducated mothers and male household head. This rule implies that 6.0 instances correctly classified while 2 instances incorrectly classified out of 8.0 instances or probably occur with success ratio of 75%.

Rules #4, 5 and 6 showed predicted under- weight of under-nutrition which was classified by J48 as IF-THEN rule and occurred in Dire Dawa, Harari and Oromia regions. A mother who visited health care services one to three times was the common predictor of under-weight malnutrition in both Dire Dawa and Harari. In the same way, uneducated mothers were common predictors of under-nutrition occurrence in Harari and Oromia. These rules implied that underweight most probably occurred with 69%, 66% and 84% success ratios within Dire Dawa, Harari and Oromia regions respectively.

Rules #7 and 8 are indicated normal nutritional status of under-five children. Mothers those completed ANC visits by attending health institution; their children more likely normal nutritional status in urban and Addis Ababa region. These rules imply that normal nutritional status occurred with success ratio of 75% in urban and 74% in Addis Ababa.

Rules #10 and 11 indicated wasted under-nutrition of under-five. Mothers those didn't visit health care institution for ANC services was the common predictors of wasted under-nutrition in Harari and Addis Ababa regions. But uneducated mothers were the predictors of wasted under-nutrition in addition to ANC visits in Addis Ababa. Rules #10 and 11 revealed that wasted under-nutrition more likely occurred with 100% success ratio in Harari and 67% in Addis Ababa.

Deployment of generated rules

Once a data mining model is built and validated, it can be used by organization (20). Among all generated rules by J48 classifier, only 11 rules were selected and explained. These rules were selected and organized by the investigator in cooperation with child nutrition coordinator of FMoH.

Now the selected rules must be organized in a way that the customer can use them in decision making. Decision such as planning, researches, setting guidelines, monitoring the planned activities and reviewing process' can be facilitated by the selected rules at national to regional level. Since this project work for the purpose of education; the investigator has tried to put generated rules in organized form.

5.4 Evaluating generated rules with knowledge discovered parameters

Data mining is a technology that uses various techniques to discover hidden knowledge from heterogeneous and distributed historical data stored in large data bases of an organization. The characteristics of newly discovered patterns are valid, novel, useful and understandable (20).

- **Valid**: not only represent current state, but also hold on new data with some certainty
- **Novel:** non-obvious to the system that are generated as new facts
- **Useful**: should be possible to act on the item or problem
- **Understandable**: humans should be able to interpret the pattern

Thus, each rule was evaluated by using measuring characteristics of discovered knowledge such as validity, useful, novel and understandable as illustrated on the tables 16.

Table: 16 Summary of evaluation results of selected rules

No	Extracted rules	Validity/certainty(5)	Useful (5)	Novelty (5)	Understandability(5)	Total mark (20)	Rank
1	If of ANC Visits = 1-3 visits AND Region=Gambela AND Child age = 36-47m AND Type of Toilet Facility=No facility/bush/field AND SOD-water =Improved AND sex of child = Female then **Stunted** (6.0/3.0)or 67%	.5	4	4	4	19	1
2	If number of ANC Visits = 1-3 visits AND Region =Tigray AND Child age = 36-47m AND birth interval=<48m AND mother age=30-34 years then **stunted** (8.0/2.0) or 80%	3	4	4	5	18	2
3	If Child age = 36-47m AND Place of delivery = Home AND Wealth-I = Poorest AND Mother Education = No AND Total member in HH = 4-6 AND Sex of house hold head = Male AND Age of house hold head = 37-42yrAND Region = Amhara: then **Stunted** nutritional	.4	4	4	3	17	3

	status occurs with success (6.0/2.0) (75%)						
4	If No of ANC visits = 1-3 visits AND Region= Dire Dawa then **underweight** (184.0/81.0) or 69%.	3	4	4	5	16	4
5	If No of ANC visits=1-3 visits AND Region =Harari AND mother education= no education then **underweight** (38.0/20) or 66%.	3	5	3	4	15	5
6	If No of ANC visits=4 visits AND mother education= no education AND Region =Oromia AND child age=48-59m AND Residence=rural then **underweight**(58.0/11.0) or 84%	4	4	3	3	14	6
7	If No of ANC visits=4 visits AND Region =Oromia AND childe age=48-59 AND residence =Urban then **normal** (6.0/2.0) or 75%	4	2	4	3	13	7
8	If No of ANC Visits= 4 visits AND region =Addis Ababa then **Normal** (35.0/12.0) or 74%	2	3	3	4	12	8
9	If No of ANC visits=no visits region =Addis Ababa and mother education= no education AND SOD water=improved then **underweight**(5.0/1.0) or 84%	4	2	2	3	11	9
10	If No of ANC visits = no visits AND Region =Harari then **wasted** (10.0) or 100%	2	3	2	3	10	10
11	If No of ANC visits = no visits AND Region= Addis Ababa AND mother education= no education then **wasted**(2.0/1.0) or 67%	1	2	3	2	8	11

As we observed from the above table all rules were ranked based on evaluating criteria.

5.5 Discussion

This project aimed to build the model for predicting under-nutrition in children less than five years at national level. The methodology employed to conduct this project work was CRISP-DM. The researcher has used J48, Naïve bayes and PART rule induction algorithms for experiments conducted on 2014 EMDHS data. Among of these algorithms J48 had best performance with accuracy; WTPR and WROC are 63.5%, 63.6% and 82.2 respectively. As a result, 11 rules were selected out of 173 rules. Similar to this work, such data mining algorithms as J48, Naïve bayes and PART rule induction were used to build predictive models on Ethiopian child nutritional status using the 2012 EDHS data. The best performing algorithm selected to build predictive models using the 2012 EDHS data was PART unpruned rule induction with 92.6% of the accurate result and 97.8% WROC area (35).

By this project work, 11 rules were selected out of 173 rules extracted by J48 algorithm using such independent attributes (predictors) as Region, place of residence, sex of child, Child age, Birth interval, Mother age, Mother education, Place of delivery, Age of household head, Sex of household head, Parity, Total member in household, Wealth index, source of drinking water, type of toilet facility, household has radio, household has TV and number of ANC visits. In addition, the class attributes (dependent variables) were stunted, underweight, wasted and normal.

The result of this project indicated that in Tigray region such attributes as child age, numbers of ANC visits, parity and mother age were identified as contributing factors for the occurrence of stunted class. Stunted children mostly found in Tigray, Amhara and Gambela. Similarly, research conducted by Assegidech in Tigray region at 2014 revealed that the significant determinants of malnutrition were child age, numbers of ANC visits, parity, mother age, education of house hold, the presence of latrine facility in the house hold, use of treated water and sex of child. Her study showed the high prevalence of stunted children 51.18% in the region with its above listed contributing factors (33).

According to generated rules by this project work; low ANC coverage, unimproved source of drinking water, unimproved type of toilet facility, delivery at home, mother education and child age are common predictors of child under-nutrition in several regions. This problem is leading to an increase in child under nutrition. These situations were similar with the report of FMoH on the annual performance of 2013/2014. From the year, 2011-2014 Child malnutrition problems have less declined in Tigray, Gambela and Amhara regions. Children in these regions, averagely 5 to 12 times suffer from diarrheal episodes in every year, resulting primarily from poor environmental sanitation such as unimproved toilet facility and unimproved source of drinking water that contributes to the child under-nutrition problems (3). Similarly, inadequate maternal care such as ANC coverage, skilled birth attendant and unhealthy environment leads to the immediate causes of under-nutrition (9).

The rule generated by this project work showed that if there are no ANC visits and a mother has no education in Addis Ababa and Harari Region then there is high probability of a child to be wasted.

CHAPTER SIX

6. CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Under-nutrition is a major public health issue in Ethiopia. To make intervention for this problem, the government has been striding for a long period of time and several researches were conducted in different parts of Ethiopia. Furthermore, data mining application is a new technology which is coming highly benefit in health care service delivery to discover hidden patterns which helps to intervene under-nutrition. Classification algorithms of data mining application applied on 2014 EMDHS to develop model. Such developed models help health care providers for better decision making such early detection and prevention of child under-nutrition problems.

The final data set identified and used for experimentation contains 5,177 instances and 19 variables including target variable. The selected target variable in this project work is under-nutritional status (2024 in stunted category, 1460 in underweight category, 1003 in normal category and 557 in wasted category). Under-nutrition is the biggest health problem of children and impossible to solve without understanding its causes. This project identified and described 19 variables which are the most contributors for the occurrence of under-nutrition. Therefore, health care providers take action on them for better reduction of under-nutrition throughout the country.

Data preprocessing activities were carried out to bring quality of output or build a model. To build predictive model, several experiments were done by J48, PART rule and Naive Bayes, but J48 was selected based on its best performance. It generated 173 rules. Among of these rules only 11 were selected. So that J48 is the best classifier for predicting under-nutrition. Each rule was evaluated by using measuring characteristics of discovered knowledge such as validity, useful, novel and understandable to the end users. After evaluation was completed, extracted rules were accepted by the owners. These extracted rules will help health care providers to make better decision by integrating them in planning and evaluating nutritional reports from national to regional.

As extracted and selected rules showed, wasted class of under-nutrition had occurred in Harari and Addis Ababa regions due to mothers did not visit health institution for ANC services. Additionally, uneducated mothers were the predictor of wasted under-nutrition in Harari. Stunted class of under-nutrition has occurred in Amhara, Gambella and Tigray regions, while Underweight under-nutrition has occurred in Oromia, Harari and Dire Dawa. By following these geographical locations, the health workers and planners easily detect and act on under-nutritional status.

Finally, to improve the intervention of under- nutrition among children less than five years; the predictors were identified, the models were built and extracted rules were summarized and presented as the owners easily understand and used in any decision making process, then integrated into planning and used for evaluating the report.

6.2 Recommendation

The investigator of this project suggested the following recommendations based on the finding.

- FMoH in a collaboration with regional health bureaus of Gambela, Tigray and Amhara regions should focus and give special care to children whose age are 36-47 months in order to manage under-nutrition problem.
- FMoH and Gambela health bureau should improve source of drinking water, Hygiene and sanitation to reduce under-nutrition problems.
- The government should strengthen PFSA in all aspects. PFSA should extend its service coverage to all health facility. Stock and consumption levels of essential nutritional products should be regularly reported to UNICEF, WHO, PFSA and FMoH from the peripheral workers.
- The FMoH and regional health bureaus should build the capacity of health workers on cutoff factors which contribute to the occurrences of under-nutrition
- The FMoH and regional health bureaus should work together to capacitate workers on proper nutritional data management as well as use of action-oriented performance monitoring at all levels.
- FMoH, Addis Ababa and Harari regions should strengthen mothers visit to health institution for ANC services in order to reduce wasted under-nutrition.
- FMoH, Harari, Oromia and Addis Ababa regions in collaboration with FMoE should improve mothers' education to reduce child under-nutrition.

Reference

1. Francis Kimani and S.K. Sharif OGW. National Guideline for Integrated Management of Acute Malnutrition, Kenya, UNICEF, 2009

2. United Nations Children's Fund (UNICEF). Improving child nutrition: the achievable imperative for global progress, New York, USA, UN, 2013

3. Federal Ministry of Health; HSDP-IV, annual performance report2013/14; Addis Ababa, Ethiopia, MOH, 2014

4. Blossner, Monika, de Onis and Mercedes. Malnutrition*:* quantifying the health impact at national and local levels*:* Environmental Burden of Disease Series, No. 12. Geneva, WHO, 2005

5. World Health Organization, Multicenter Growth Reference studies Group (2006). WHO child growth Standards*:* Length/height-for-Age, Weight-for-Age, Weight-for-height and Body mass Index-for-Age*:* Methods and Development. Geneva, Switzerland, WHO, 2006

6. Federal Democratic Republic of Ethiopia, MOH. Ethiopian National Nutrition Programme June, 2013-June 2015.Addis Ababa, Ethiopia, MOH, 2013

7. Federal Democratic Republic of Ethiopia. National monitoring and reporting system for the implementation of community-led total sanitation and hygiene**;** Addis Ababa, Ethiopia, MOH, 2012

8. Federal Democratic Republic of Ethiopia, MOH. Health Sector Development Program IV, 2010/2011-2014/15.Addis Ababa, Ethiopia, MOH, 2010

9. Dr. Pat Pridmore and Professor Roy Carr Hill. Addressing the underlying and basic causes of child under nutrition in developing countries: what works and why? India, Denmark, 2009.

10. Central Statistical Agency [Ethiopia]. Ethiopia Mini Demographic and Health Survey2014/2015, Addis Ababa, Ethiopia, 2015

11. Todd Benson , Abera Kumie, Solomon Bellete , Demese Chanyalew , Tefera Belachew , Ayele Gebremariam, Damene Hailemariam, Eleonora Genovese & Fikru Tesfaye. An assessment of the causes of malnutrition in Ethiopia: A contribution to the formulation of a National Nutrition Strategy for Ethiopia. International Food Policy Research Institute, Washington, DC, USA, 2005

12. Ethiopia Central Statistics Agency and ICF International. Ethiopia Demographic and Health Survey (2011), Addis Ababa, Ethiopia, and Calverton, Maryland, USA: Central Statistical Agency and ORC Macro ,2012

13. Federal Democratic republic of Ethiopia, Ministry of Finance and Economic Development (MoFED). Growth and Transformation Plan (GTP), 2010/11-2014/15. Addis Ababa, Ethiopia, Sept, 2010

14. Kimberly Moore Waggoner (2011).evaluating the impact of water and sanitation quality on child malnutrition in sub-Saharan Africa. Thesis article, Washington, DC, School of Arts and Sciences of Georgetown University, April, 2011

15. World Health Organization (WHO) and United Nations Children's Fund (UNICEF). Joint Monitoring Program for water Supply and Sanitation. Progress on Sanitation and drinking water: 2014 Update. Geneva and New York: WHO and UNICEF, 2014

16. United Nations Children's Fund (UNICEF). Tracking progress on child and maternal nutrition: A survival and development priority; Geneva and New York: WHO and UNICEF November, 2009

17. Boris Milovic. Prediction and Decision Making in Health Care using Data Mining, International Journal of Public Health Science (IJPHS), University, Serbia Vol. 1, No. 2, pp. 69~78 ISSN: 2252-, December 2012

18. Licentiate. Predictive Techniques and Methods for Decision Support in Situations with Poor Data Quality; Rikard könig technology university of boras, school of business and informatics, 2009

19. Sumathi, Sivanandam SN. Introduction to Data: Mining and its applications. Berlin, German: Springer-Verlag inc, 2006

20. Cios Krzysztof J, Pedrycz Witold, Swiniarsk Roman W, Kurgan Lukasz A. Data mining: knowledge Discovery Approach. New York, USA: Springe Science Business Media LLC, 2007.

21. http://www.CRISP-DM.ORG

22. Two Crows Corporation. Introduction to Data mining and Knowledge Discovery; Third Edition.USA: Two Crows Corporation, 2005

23. Divya. T and Sonali. A. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266

24. Arun George Eapen. Application of Data mining in Medical Applications; University of Waterloo Systems Design Engineering Waterloo, Ontario, Canada, 2004

25. Riccardo. B, Blaz.Z. Predictive data mining in clinical medicine: Current issues and guidelines. Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX, US ,international journal of medical informatics 77 (2 0 0 8) 81–97

26. Hsinchun .C, Sherrilynne .S . Fuller Carol .F William .H. Medical informatics: Knowledge Management and Data Mining in Biomedicine.USA, Springer Science Business Media, Inc. 2005

27. Larose Daniel T. Discovering Knowledge in Data-An Introduction to Data mining. New Jersy, USA: John Wiley & Sons Inc; 2005

28. Bramer Max. Principles of Data Mining. London, Springe- Verlag Limited; 2007

29. Witten Ian H, Frank Eibe. Data Mining: Practical Machine Learning Tools and Techniques, Second edition. USA: Elsevier Inc, 2005

30. Han Jiawie, kamber Micheline. Data mining: Concepts and techniques. 3rd edition, New York. USA: Morgan Kaufmann publisher, 2012

31. Remco R. Bouckaert. E, Irank Mark.H, Richard. K Peter.R. WEKA Manual for Version 3.7.10, University of Waikato, Hamilton, New Zealand, July 31, 2013.

32. Tadiwos Zewdie and Degnet Abebaw. Determinants of Child Malnutrition: Empirical Evidence from Kombolcha District of Eastern Hararghe Zone, Ethiopia, Ethiopian Economics Association/Ethiopian Economic Policy Research Institute (EEA/EEPRI), Addis Ababa, Ethiopia,2013

33. Mengistu K, Alemu K, Destaw B, Prevalence of Malnutrition and Associated Factors Among Children Aged 6-59 Months at Hidabu Abote District, North Shewa, Oromia Regional State, research article, Institute of Public Health, College of Medicine and Health Sciences, Gonder University Ethiopia, 2013

34. Asegedech, determinants of child malnutrition a case study in central zone of Tigray, northern Ethiopia: Research article, department of economics, college of business and economics, Mekelle, Ethiopia, 2014

35. Zenebe Markos ,Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: The Case of 2011 Ethiopian Demographic and Health Survey, Research article, Volume 5 • Issue 2 • 1000152,Addis Ababa, Ethiopia, June 2013

Appendix

Appendix: A, Evaluation form for selected rules (Discovered knowledge)

No	Extracted rules	Validity/certainty(5)	Useful (5)	Novelty (5)	Understandability(5)	Total mark (20)	Rank

Appendix: B J48 classifier out put

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.1 -M 2
Relation:     children's nutritional status data of 2014EMDHS-
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    5177
Attributes:   19
              Region
              Residence
              sex of child
              Child age
              Birth interval
              Moth-Age
              Moth-Educ
              P-delivery
              Age of HH-head
              Sex of HH-Head
              Parity
              Total member in HH
              Wealth-I
              SOD-water
              TT-Facility
              HH-has-Radio
              HH-has-TV
              No ANC Visits
              NUTSTAT
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

No ANC Visits = 1-3visits
|   Region = Oromiya: Stunting (356.0/171.0)
|   Region = Dire Dawa: Stunting (126.0/43.0)
|   Region = Addis Ababa: Stunting (76.0/29.0)
|   Region = Amhara
|   |   Child age = 0-6m: Normal (1.0)
|   |   Child age = 24-35m: Underweight (68.0/27.0)
|   |   Child age = 18-23m: Underweight (29.0/10.0)
|   |   Child age = 48-59m: Stunting (29.0/9.0)
|   |   Child age = 36-47m
|   |   |   HH-has-TV = yes: Underweight (8.0/3.0)
|   |   |   HH-has-TV = No: Stunting (87.0/52.0)
|   |   Child age = 12-17m: Underweight (9.0/3.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Underweight (0.0)
|   Region = Benishangul-Gumuz: Underweight (120.0/37.0)
|   Region = Harari: Underweight (110.0/59.0)
|   Region = Gambela
|   |   Child age = 0-6m: Normal (1.0)
```

```
|   |   Child age = 24-35m: Underweight (40.0/16.0)
|   |   Child age = 18-23m: Underweight (4.0/1.0)
|   |   Child age = 48-59m
|   |   |   P-delivery = health facility: Underweight (2.0)
|   |   |   P-delivery = Home/else where: Stunting (17.0/7.0)
|   |   Child age = 36-47m
|   |   |   TT-Facility = improved: Normal (6.0)
|   |   |   TT-Facility = No facility/bush/field
|   |   |   |   SOD-water = Improved: Underweight (7.0/2.0)
|   |   |   |   SOD-water = Unimproved
|   |   |   |   |   sex of child = Male: Normal (7.0)
|   |   |   |   |   sex of child = Female: Stunting (6.0/3.0)
|   |   |   |   SOD-water = Other: Normal (0.0)
|   |   |   TT-Facility = Unimproved: Underweight (37.0/18.0)
|   |   |   TT-Facility = Improved: Underweight (0.0)
|   |   Child age = 12-17m: Underweight (12.0/1.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Underweight (0.0)
|   Region = Somali
|   |   Child age = 0-6m: Normal (3.0)
|   |   Child age = 24-35m: Underweight (78.0/25.0)
|   |   Child age = 18-23m: Underweight (16.0/1.0)
|   |   Child age = 48-59m: Stunting (40.0/9.0)
|   |   Child age = 36-47m
|   |   |   Wealth-I = Richest: Normal (11.0/6.0)
|   |   |   Wealth-I = Poorest: Stunting (20.0/9.0)
|   |   |   Wealth-I = Poorer: Stunting (15.0/5.0)
|   |   |   Wealth-I = Middle: Underweight (19.0/6.0)
|   |   |   Wealth-I = Richer: Underweight (17.0/8.0)
|   |   Child age = 12-17m: Underweight (18.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Underweight (0.0)
|   Region = Affar
|   |   Child age = 0-6m: Normal (1.0)
|   |   Child age = 24-35m: Underweight (49.0/5.0)
|   |   Child age = 18-23m: Underweight (13.0/1.0)
|   |   Child age = 48-59m: Stunting (51.0/16.0)
|   |   Child age = 36-47m
|   |   |   Age of HH-head = 37-42yr
|   |   |   |   P-delivery = health facility: Stunting (2.0)
|   |   |   |   P-delivery = Home/else where: Underweight (15.0/3.0)
|   |   |   Age of HH-head = 49-54yr: Stunting (4.0/1.0)
|   |   |   Age of HH-head = above 64yr: Normal (3.0)
|   |   |   Age of HH-head = 19-24yr: Underweight (3.0)
|   |   |   Age of HH-head = 25-30yr: Underweight (10.0/1.0)
|   |   |   Age of HH-head = 31-36yr
|   |   |   |   TT-Facility = improved: Underweight (0.0)
|   |   |   |   TT-Facility = No facility/bush/field: Underweight
(3.0)
|   |   |   |   TT-Facility = Unimproved: Stunting (9.0/3.0)
|   |   |   |   TT-Facility = Improved: Underweight (1.0)
|   |   |   Age of HH-head = 55-60yr: Normal (3.0)
|   |   |   Age of HH-head = 61-64yr: Underweight (0.0)
|   |   |   Age of HH-head = 43-48yr
```

```
|   |   |   |   HH-has-Radio = yes: Underweight (2.0)
|   |   |   |   HH-has-Radio = No: Stunting (2.0/1.0)
|   |   |   Age of HH-head = <18yr: Underweight (0.0)
|   |   Child age = 12-17m: Underweight (22.0/4.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Underweight (0.0)
|   Region = Tigray
|   |   Child age = 0-6m: Normal (1.0)
|   |   Child age = 24-35m
|   |   |   Birth interval = <24month: Normal (20.0/8.0)
|   |   |   Birth interval = >48month: Normal (11.0/3.0)
|   |   |   Birth interval = 25-47month
|   |   |   |   TT-Facility = improved: Underweight (3.0)
|   |   |   |   TT-Facility = No facility/bush/field: Underweight
(12.0/2.0)
|   |   |   |   TT-Facility = Unimproved: Normal (4.0)
|   |   |   |   TT-Facility = Improved: Underweight (0.0)
|   |   Child age = 18-23m
|   |   |   Wealth-I = Richest: Normal (3.0)
|   |   |   Wealth-I = Poorest: Normal (3.0)
|   |   |   Wealth-I = Poorer: Underweight (7.0/2.0)
|   |   |   Wealth-I = Middle: Underweight (2.0)
|   |   |   Wealth-I = Richer: Underweight (7.0/2.0)
|   |   Child age = 48-59m: Stunting (24.0/3.0)
|   |   Child age = 36-47m
|   |   |   Birth interval = <24month: Underweight (15.0/6.0)
|   |   |   Birth interval = >48month
|   |   |   |   Moth-Age = 30-34: Stunting (8.0/2.0)
|   |   |   |   Moth-Age = 15-19: Underweight (0.0)
|   |   |   |   Moth-Age = 20-24: Underweight (1.0)
|   |   |   |   Moth-Age = 25-29: Underweight (7.0/3.0)
|   |   |   |   Moth-Age = 35-39: Underweight (5.0/2.0)
|   |   |   |   Moth-Age = 40-44: Underweight (0.0)
|   |   |   |   Moth-Age = 45-49: Underweight (1.0)
|   |   |   Birth interval = 25-47month
|   |   |   |   Residence = Rural: Normal (33.0/17.0)
|   |   |   |   Residence = Urban: Stunting (3.0)
|   |   Child age = 12-17m: Underweight (8.0/1.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Underweight (0.0)
|   Region = SNNP: Stunting (326.0/90.0)
No ANC Visits = no visits
|   Region = Oromiya: Normal (152.0/67.0)
|   Region = Dire Dawa: Stunting (65.0)
|   Region = Addis Ababa
|   |   Moth-Educ = Primary: Stunting (3.0)
|   |   Moth-Educ = Secondary: Wasted (2.0/1.0)
|   |   Moth-Educ = No education
|   |   |   SOD-water = Improved: Stunting (3.0)
|   |   |   SOD-water = Unimproved: Underweight (5.0/1.0)
|   |   |   SOD-water = Other: Underweight (0.0)
|   |   Moth-Educ = Higher: Stunting (0.0)
|   Region = Amhara: Stunting (248.0/82.0)
|   Region = Benishangul-Gumuz: Stunting (205.0/62.0)
```

```
|   Region = Harari: Wasted (10.0)
|   Region = Gambela: Stunting (176.0/58.0)
|   Region = Somali: Stunting (278.0/84.0)
|   Region = Affar: Stunting (269.0/85.0)
|   Region = Tigray: Stunting (253.0/85.0)
|   Region = SNNP: Normal (25.0)
No ANC Visits = above 4 visits
|   Region = Oromiya
|   |   Child age = 0-6m: Normal (2.0)
|   |   Child age = 24-35m: Normal (1.0)
|   |   Child age = 18-23m: Normal (3.0)
|   |   Child age = 48-59m
|   |   |   Residence = Rural: Underweight (58.0/11.0)
|   |   |   Residence = Urban: Normal (6.0/2.0)
|   |   Child age = 36-47m: Normal (5.0)
|   |   Child age = 12-17m: Normal (5.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Normal (2.0)
|   Region = Dire Dawa: Underweight (184.0/81.0)
|   Region = Addis Ababa: Normal (35.0/12.0)
|   Region = Amhara
|   |   Child age = 0-6m: Normal (2.0)
|   |   Child age = 24-35m: Underweight (0.0)
|   |   Child age = 18-23m: Normal (4.0)
|   |   Child age = 48-59m: Underweight (74.0/33.0)
|   |   Child age = 36-47m: Normal (6.0)
|   |   Child age = 12-17m: Normal (3.0)
|   |   Child age = 6-8m: Underweight (0.0)
|   |   Child age = 9-11m: Normal (3.0)
|   Region = Benishangul-Gumuz: Normal (89.0/22.0)
|   Region = Harari
|   |   Child age = 0-6m: Stunting (9.0/2.0)
|   |   Child age = 24-35m: Normal (34.0/15.0)
|   |   Child age = 18-23m
|   |   |   Moth-Educ = Primary: Normal (7.0/2.0)
|   |   |   Moth-Educ = Secondary: Stunting (1.0)
|   |   |   Moth-Educ = No education: Stunting (3.0)
|   |   |   Moth-Educ = Higher: Stunting (0.0)
|   |   Child age = 48-59m
|   |   |   Moth-Educ = Primary
|   |   |   |   P-delivery = health facility: Underweight (2.0/1.0)
|   |   |   |   P-delivery = Home/else where
|   |   |   |   |   sex of child = Male: Normal (5.0)
|   |   |   |   |   sex of child = Female: Stunting (5.0/1.0)
|   |   |   Moth-Educ = Secondary: Underweight (0.0)
|   |   |   Moth-Educ = No education: Underweight (38.0/20.0)
|   |   |   Moth-Educ = Higher: Underweight (0.0)
|   |   Child age = 36-47m: Normal (21.0/7.0)
|   |   Child age = 12-17m: Stunting (8.0/3.0)
|   |   Child age = 6-8m: Normal (0.0)
|   |   Child age = 9-11m: Stunting (6.0/2.0)
|   Region = Gambela
|   |   Moth-Educ = Primary: Normal (7.0/2.0)
|   |   Moth-Educ = Secondary: Normal (2.0)
```

```
|   |   Moth-Educ = No education
|   |   |   Child age = 0-6m: Underweight (0.0)
|   |   |   Child age = 24-35m: Underweight (0.0)
|   |   |   Child age = 18-23m: Underweight (4.0/2.0)
|   |   |   Child age = 48-59m: Underweight (51.0/22.0)
|   |   |   Child age = 36-47m: Underweight (1.0)
|   |   |   Child age = 12-17m
|   |   |   |   HH-has-Radio = yes: Underweight (2.0)
|   |   |   |   HH-has-Radio = No: Normal (5.0)
|   |   |   Child age = 6-8m: Underweight (0.0)
|   |   |   Child age = 9-11m: Normal (2.0)
|   |   Moth-Educ = Higher: Normal (1.0)
|   Region = Somali
|   |   Child age = 0-6m: Normal (5.0)
|   |   Child age = 24-35m: Normal (23.0)
|   |   Child age = 18-23m: Normal (8.0)
|   |   Child age = 48-59m: Underweight (89.0/38.0)
|   |   Child age = 36-47m: Normal (21.0/1.0)
|   |   Child age = 12-17m: Normal (16.0)
|   |   Child age = 6-8m: Normal (0.0)
|   |   Child age = 9-11m: Normal (3.0)
|   Region = Affar: Normal (181.0/44.0)
|   Region = Tigray
|   |   Child age = 0-6m: Normal (1.0)
|   |   Child age = 24-35m: Underweight (0.0)
|   |   Child age = 18-23m: Normal (7.0)
|   |   Child age = 48-59m
|   |   |   HH-has-TV = yes: Stunting (2.0/1.0)
|   |   |   HH-has-TV = No: Underweight (56.0/19.0)
|   |   Child age = 36-47m: Normal (5.0)
|   |   Child age = 12-17m: Normal (9.0)
|   |   Child age = 6-8m: Normal (1.0)
|   |   Child age = 9-11m: Normal (1.0)
|   Region = SNNP: Underweight (267.0/108.0)

Number of Leaves  :    173

Size of the tree :     214
Time taken to build model: 0.61 seconds

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

Correctly Classified Instances        3290               63.5503 %
Incorrectly Classified Instances      1887               36.4497 %
Kappa statistic                          0.4572
Mean absolute error                      0.2428
Root mean squared error                  0.3561
Relative absolute error                 68.5905 %
Root relative squared error             84.6391 %
Coverage of cases (0.95 level)          97.3537 %
Mean rel. region size (0.95 level)      68.8768 %
Total Number of Instances             5177
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.018	0.002	0.500	0.018	0.035	0.079	0.732	0.224	Wasted
	0.656	0.178	0.591	0.656	0.622	0.464	0.812	0.597	Underweigh
	0.872	0.301	0.651	0.872	0.745	0.558	0.831	0.690	Stunting
	0.491	0.066	0.676	0.491	0.569	0.480	0.864	0.657	Normal
Av/w	0.636	0.183	0.623	0.636	0.595	0.463	0.822	0.606	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
10	102	425	20	a = Wasted
5	958	361	137	b = Underweight
4	145	1765	110	c = Stunting
1	415	162	557	d = Normal