2019-09

Using data mining to predict the status of anemia among children aged 6 - 59 months

Gebremariam, Adino

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

COMPUTING FACULTY

PREDICTING THE STATUS OF ANEMIA AMONG CHILDREN AGED 6 - 59 MONTHS USING DATA MINING TECHNIQUES

By

ADINO GEBREMARIAM

BAHIRDAR, ETHIOPIA
September, 2019

USING DATA MINING TO PREDICT THE STATUS OF ANEMIA AMONG CHILDREN AGED 6 - 59 MONTHS.

By
Adino Gebremariam

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology, BDU in partial fulfillment of the requirements for the degree

Of

Master of Science in Information Technology

Advisor: Gebeyehu Belay (Dr of Eng.) Asct Prof.

Bahir Dar, Ethiopia

September 2019

DECLARATION

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

Name of the student: Adino Gebremariam

Signature_____

Date of submission: _____

Place: Bahir Dar

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

Advisor Name: Gebeyehu Belay (Dr of Eng.) Asct Prof.

Advisor's Signature: _____

BY : ADINO G/MARIAM EWUNU

APPROVAL SHEET

Name and Signature of Members of the Examining Board

Belete B.		26/02/2012 E.C
Faculty Dean	Signature	Date
Gebeyehu B (Dr)		
Advisor	Signature	Date
Tibebe B.		Oct 28, 2019
External Examiner	Signature	Date
Sefi G.		06/11/2019
Internal Examiner	Signature	Date

DEDICATION

This work must be dedicated to my beloved families.

ACKNOWLEDGEMENT

First and foremost extraordinary thanks go for my Almighty God and His Mother Saint-Merry, for providing me this opportunity granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

Dr. Gebeyehu Belay, my respected Advisor, my cordial thanks for giving me the freedom to do my thesis in this area, for your warm encouragement, thoughtful guidance, critical comments, and correction of the thesis. I would like to thank you especially for friendly assistance with various problems all the time.

I also wish to extend my sincere gratitude to all my instructors who have been the source of all my achievements. Furthermore, I would like to express my sincere gratitude to Computer Science and IT PG coordinator Mr. Asegahegn Endalew for all his enthusiasm and continuous support.

Finally, I would like to express my special heartfelt thanks go to my parents, brothers, sisters, my wife and to my son for their understanding and provision of unrestricted support and encouragement through both the ups and downs of my time during my study.

ABSTRACT

Anemia is recognized as major public health problems globally with diverse consequences for human health as well as socioeconomic development (WHO, 2015). Globally, more than two-fifths of young children in the world are affected by anemia. In Ethiopia, based on the 2005, 2011, and 2016 EDHS, 54%, 44% and 57 of children aged 6 to 59 months are anemic, respectively. The aim of this study is to develop a model for predict the status of anemia among children aged 6 to 59 months using data mining techniques. This study followed hybrid methodology of Knowledge Discovery Process to achieve the goal of building a predictive model using data mining techniques and used secondary data from the 2016 Ethiopia Demographic and Health Survey (EDHS) dataset of 8603 records of both anemic and not anemic children aged 6 to 59 months through five experiments and ten scenarios. WEKA 3.9.3 data mining tools and classification techniques such as J48 decision tree, Naïve Bayes and PART rule induction algorithms were employed as means to address the research problem. Model comparison is done based on TP (sensitivity) and FP (specificity) rates, precision, recall, F-measure, ROC area and accuracy. 10-fold cross validation test option is used to check the performances of each classifier. In this particular study, the predictive model developed using J48 pruned with all attributes perform better in predicting anemic cases with an accuracy of 91.805%. Generally the results from this study were encouraging and confirmed that applying data mining techniques could indeed support a predictive model building task that predicts anemic status children aged 6 to 59 months in Ethiopia. Thus, the outcome of this study helps health care planners and policy makers to design a proper and suitable preventive and control program to combat anemia. In the future, integrating large demographic and health survey dataset and clinical dataset, employing other classification algorithms, tools and techniques could yield better results.

Keywords: EDHS Dataset, Anemia, Data Mining, Predictive Model, Classification, J48, Naïve Bayes and PART rule, WEKA.

TABLE OF CONTENTS

List of Acronyms

ANC	Antenatal Care
ARFF	Attribute Relation File Format
BMI	Body Mass Index
CRISP-DM	Cross-Industry Standard Process For Data Mining
CSA	Centeral Statics Agency
CSV	Comma Separated Value
DHS	Demographic And Health Surveys
DM	Data Mining
DT	Decision Tree
EDHS	Ethiopia Demographic And Health Surveys
FN	False Negative
FP	False Positive
g/dl	Gram Per Deci Litre
g/l	Gram Per Litre
KDD	Knowledge Discovery Databases
NB	Naïve Bayes
PNC	Postnatal Care
ROC	Receiver Operating Characteristics
TN	True Negative
TP	True Positive
WEKA	Waikato Environment For Knowledge Learning
WHO	World Health Organization

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

1. INTRODUCTION

1.1. BACKGROUND

Anemia is one of the most frequently observed nutritional deficiency diseases in the world today. It is a disorder characterized by the concentration of hemoglobin in the blood lower than the defined normal level, and it is usually related to a decrease in the circulating mass of red blood cells (WHO, 1992).

Anemia is one of the major public health problems globally with diverse consequences for human health as well as socioeconomic development (WHO, 2015). It affects the physical health and cognitive development of individual causing low productivity and poor economic development of a country (Stevens GA et al. 2013). The problem is also related to high maternal and child morbidity and mortality, especially in developing countries (WHO, 2014). It is the most common disorder of the blood affecting more than 2 billion people globally, accounting for over 30% of the world's population (WHO, 2015). It is the most common public health problem in all types of people, particularly in low income countries. Pregnant women and young children carry the highest burden (Rodak and Bernadette, 2007).

Anemia in children has been associated with impaired cognitive performance, motor and language development, and scholastic achievement (WHO, 2008). In Ethiopia, 57% of children age 6-59 months suffered from some degree of anemia (hemoglobin levels below 11 g/dl) such as mild anemia (10-10.9), moderate anemia (7-9.9), severe anemia (<7) (EDHS, 2016).

Hemoglobin is the protein molecule in red blood cells that carries oxygen from the lungs to the tissues of the body and returns carbon dioxide from the tissues back to the lungs (charles Patrick Davis,1996).

In Ethiopia, so far four Demographic and Health Surveys (EDHS) have been conducted at five-year intervals since 2000 with the primary objectives of providing quality information for planning, policy formulation, monitoring and evaluation of population, and health programs in the country.

According to EDHS (2016), the primary purpose of conducting the DHS in Ethiopia was to furnish policy-makers, planners, researchers, and program managers with detailed information of the population and health situation, covering topics on family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, malaria, anemia, woman's empowerment, and knowledge of HIV/AIDS.

Faced with the tremendous economic and competitive pressures, the health-care industry has started to mine its data to minimize costs, enhance quality and save lives. In support of this notion, (Bresnahan J, 2010) argued that one way in which data mining is helping health-care providers cut costs and improve care is by showing which treatments have been most effective. The data mining process which serves as a means of searching previously unknown, actionable information from large databases can be used to improve the quality, efficiency and care of patients which is known in the health care industry as outcomes measurements.

To predict the status of anemia the researcher uses the data mining techniques. Data mining is the process of analyzing data from different perspectives and summarizing the results as useful information. Data mining, in the health care sector, is set to play an important role in tackling the data overload (Han J, Kamber, M, 2006). In this regard, (Daniel T. L, 2005) states that benefits of data mining include improving health care

quality reduced operating costs, and better insight into health data. Given the benefits acquired in applying data mining techniques in the health care sector, it is proper to assess the relevance and potential advantage of such emerging technology in the Ethiopian health sector.

Thus, since there is still a moderate prevalence of anemia in Ethiopia, in this study, the researcher is motivated to explore the potential applicability of data mining techniques to predict the status of anemia children aged 6 to 59 months in Ethiopia by using the Ethiopia Demographic and Health Survey of 2016 datasets. And also, this study investigated the major risk factors of anemia, which will help to guide health professionals, program managers and health policy makers to identify indicators for monitoring anemia strategy and applying necessary preventive and appropriate measures to decrease anemia disease using data mining techniques.

1.2. Statement of the problem

The reason that initiated this research work is the high rate of anemia disease in Ethiopia according to the (EDHS, 2016) and it affects the physical health, cognitive performance, motor and language development, and scholastic achievements of children.

Anemia is recognized as a main public health problem throughout the world (Sharman A, 2000). And occurs at all stages of the life cycle, but is more dominant in pregnant women and young children (WHO, 2008). According to the epidemiological data collected from multiple countries of DHS by the World Health Organization (WHO), more than one-third of women and two-fifths of young children in the world are affected by anemia (WHO, 1992). In Ethiopia, 57% of children age 6-59 months suffered from some degree of anemia (hemoglobin levels below 11 g/dl) and huge amount of data was available in the central static agency (EDHS, 2016). As the researcher observed the EDHS website, the amount of data stored in the dataset is huge. However, these databases can be used to exploit the hidden information behind the problems of anemia and is important to take

remedial actions. According to the EDHS (2016) between 2005 and 2011, the prevalence of anemia among Ethiopian children declined from 54% to 44% from 2005 to 2011, but increased to 57% in 2016.

Analyzing the past occurrence of anemia who already took the causes of anemia would provide a better perspective of the reasonable solutions to predict the status of anemia. This can be achieved using the concepts of data mining.

Model with high accuracy is beneficial for predicting the status of anemia among children aged 6 to 59 months. It is required that health sector, families and other stakeholders work on the identified factors, so that anemia disease could be minimized in the future. Therefore, the need for predicting the status of anemia in this research work is to determine the reasons and work effectively to solve the problems of anemia. The aim of the research is therefore to construct anemia predictive model using data mining techniques so as to predict the status of anemia child aged 6 to 59 months using EDHS database. There is also an apparent lack of a model in the developing world focused towards to predict the status of anemia children aged 6 to 59 months. This thesis will be carried out to fill this gap, to this end, the research attempts to answer the following questions:

- What are the factors that influence the occurrence of anemia in children aged 6 to 59 months?
- What are the potential sets of attributes to develop a model to reasonably predict any anemic children?
- Which classification algorithm can be more suitable for the purpose of predicting anemia among children aged 6 to 59 months?

1.3. **Objectives**

1.3.1. General objective: -

The general objective of this study is to predict the status of anemia among children aged 6 - 59 months using data mining techniques

1.3.2. Specific objectives:

To achieve the general objective indicated above, the researcher might be accomplished the following specific objectives:

- ♣ To identify the factors that influences the occurrence of anemia in children aged 6 to 59 months.
- ♣ To compare the models based on their classification accuracy and performances measures and select the best classification model.
- ♣ To build predictive models in order to classify children to any anemic/not anemic levels.
- ♣ To determine the potential set of attributes which can reasonably predict any anemic children aged 6 to 59 months.

1.4. Scope and limitation of the study

The study mainly focuses on developing a predictive model that is capable of predicting anemic cases in children aged 6 to 59 months in Ethiopia using data mining tools and techniques, specifically classification techniques. It also covers all regions and the two city administrations (Addis Ababa and Dire Dawa) of Ethiopia. The scope of this research is also restricted to utilizing the secondary data of the 2016 Ethiopian Demographic and Health Survey (EDHS) dataset collected by CSA.

The unavailability of related literature, particularly in relation to the application of data mining technology on epidemiological datasets is one of the limitations encountered to

undertake the study. And also, while conducting the study, a number of difficulties might be encountered, including shortage of sufficient finance, time and difficulty to filtering anemia data from the whole EDHS data.

1.5. Significance of the study

The outcome of this research has been a great contribution for different stakeholders/ various areas for different decision-making purposes. Further it will have an advantage to show the influential factors which anemia disease and help in giving timely managerial decisions.

These studies show the correlation or association of the different factors of socioeconomic, geographic, maternal health, biological factor etc., with the status of anemia. And also, the study will be important for public health planners to plan and implement preventive actions focused on child health strategies so as to mitigate the risk of anemia attributed by avoidable recommended factors.

Regional health office may use the data mining technique to determine the level of the factors behind anemia among children aged 6 to 59 months and may recommend and provide necessary information for the health sector and stakeholders, health professional, families and other concerned bodies to realize the magnitude of the problem so as to enable them to make immediate remedial actions on the identified factors. This helps the organization to design and develop further strategies for the efforts towards solving the problems of anemia. Health office may also identify its potential, focus on the most identified variable for anemia. On the other direction of the study will be useful for further consultancy for research.

1.6. Thesis Organization

The thesis is organized into five chapters; the first chapter is an introduction part, which contains background to the research work, and related concepts, statement of the problem, objectives, scope and limitation of the study and the significance of the study.

The second chapter reviews data mining as a technology and techniques & applications, health data mining, and explaining about literature review on data mining, methods/techniques used, and its application in the health care sector and reviewing of related works.

The third chapter explains the methodologies adopted for the purpose of this research work. To give an understanding about the data mining tool, techniques and algorithms that applied in the study and issues related to data analysis tool, classification technique and model evaluation techniques.

The fourth chapter is discussed about on data preprocessing tasks including data cleaning, transformation and attribute selection. And also presents about the results of the analysis and the rules discovered and evaluation of the discovered knowledge. Detailed experimentation and analysis was also discussed at this chapter. The last chapter, chapter five, will provide conclusion, and offers recommendations based on the research findings.

CHAPTER TWO

LITERATURE REVIEW

2.1. Overview of Data Mining

Development in digital data acquisition and storage technology has resulted in the growth of huge databases. This can be seen in different sectors; for instance, supermarket transactional data, telephone call details, different governmental statistics, credit card records, different medical records. These days interest has grown in the possibility of extracting information from the databases that might be of valuable to the owner of the database. The discipline concerned with the task has become known as Data Mining (Deogan, 2011).

Also, the tools and techniques for data collecting, storing and transferring for different purposes have also increased. However, this massive volume of stored data needs to be extracted and suitable for gaining information and knowledge; unless they are no value without extracting efficiently in order to get information from them (F. T, 2006). This raises the demand of new tools and techniques which is helpful in analyzing the massive data for information and knowledge. This leads to the idea of data mining; Data mining is often set in the broader context of knowledge discovery in data bases, or KDD. According to (R. Agrawal, 2015) the KDD process contains several stages; selecting the target data, preprocessing the data, transform the data, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures.
It is estimated that the amount of data stored in the world's database raises every twenty months at a rate of 100% (Witten and Frank, 2000).

As the size of data rises, the proportion of information in which people could understand decreases considerably. This tells that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which

results with increasing information gap. Consequently, people begin to realize this bottleneck and to look into possible remedies.

2.1.1. What is data mining?

Data Mining is a concept which has different meaning regarding different researchers. According to T.Sutch, (2015), Data Mining is the process of analyzing data from different views and summarizing the results as useful information.

According to Fayyad and Piatetsky, (2010), data mining is a process of non-trivial extraction of implicit, previously unknown, potentially useful and actionable information (such as knowledge rules, constraints, regularities) from huge amounts of data in databases. This information enables to make serious business decision.

Data mining is the computing processes of discovering patterns, hidden information and unknown data, relationships and knowledge in large datasets. It is a confluence to machine learning, statistics, Artificial Intelligence, database and others. It requires to analyze large-scale data, which cannot handle by traditional statistical methods (Jiawei al et, 2012). It is an important process where intelligent methods are applied to extract patterns and is a key step in the overall process of knowledge discovery in databases (Mary & Obenshain, 2004).

2.3. Data mining techniques in the Study

According to A.Olani, (2008), the objective of data mining is both predictions which have an aim of predicting unknown or future values of the attributes of interest using other attributes in the databases and description which has an aim of describing the data in a manner understandable and interpretable to humans. The purpose of data mining is either to create a descriptive model or a predicative model. The relative importance of description and prediction depend on the use in different applications. These two goals can be achieved by any of a number of data mining tasks including: classification,

regression, clustering**,** summarization, association dependency modeling, and deviation detection.

2.3.1. Predictive modeling

As stated in (P. Neelamadhab, et al 2012), predicative model permits the value of one variable to be predicted from the known values of other variables. Predicative data mining methods predicts the values of data, using some already known results that have been found using a different set of data. Predicative data mining tasks include; Classification, Prediction, Regression, and Time Series Analysis (S. Fadzilah and A. Mansour, 2011).

In this study, the researcher was implemented classification data mining technique with three different algorithms such as J48, Naïve Bayes and PART rule induction due to the nature of the data and the objective of the study.

Classification is the most commonly applied data mining technique, for the development of predictive models on pre-classified instances. In the next sub section classification data mining approach is discussed which is applied in the data in the conducted research.

2.3.1.1. Classification

Classification is a supervised learning i.e. there are pre-specified target variables. During classification there is a mapping from some input variable to a categorical variable. The input variables of from training data set are used to build a model that classifies new data into class labels. The most commonly used techniques of classification are Artificial Neural Network, Decision tree, and Naïve Bayes etc., (Guo Y, Grossman R., 1999).
Different algorithms are applied to achieve the tasks of data mining, common algorithms used for classification are; Artificial neural network, decision tree and Naïve Bayes.

According to Bharati (2013), Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples.

2.4. Data mining process models

The data mining process models can be considered as a methodology to support the process which leads to find the information and knowledge. The reason for using the process models is in order to organize the knowledge discovery and data mining projects within a common frame work. Also, the process models are helpful to understand the knowledge discovery process and provide a roadmap while planning and carrying out the projects (R. Pressman, 2005).

Moreover ,the reasons of using the process models which are mentioned in a research study that to ensure the end product will be useful for the users (T. Sutch, 2015), the other reason which is pointed out by (T. Sutch, 2015) is that to understand the process itself and to understand the concerns and need of the end users. End users usually lack perception of large amounts of untapped and potentially valuable data. Besides they are not ready to devote time and resources toward formal methods of knowledge seeking. Another reason of a need of data mining process models is mentioned by (R. Pressman, 2006) is that providing support for managerial processes.

There are different classes of data mining process models. Each model has its own strength and weakness. KDD process (Knowledge Discovery in Databases), CRISP-DM

(Cross Industry Standard Process for Data Mining), SEMMA (Sample Modify Model Assess), and hybrid model are some of the models that are used in different DM projects.

For this study hybrid data mining process model is used, which combines both KDD and CRISP-DM features. The selected process model has high importance for the selected research domain.

2.4.1. The KDD Process Model

As stated by Cao. L (2007) Knowledge-Discovery and Data Mining (KDD), is the process of automatically searching large volumes of data for hidden, interesting, unknown and potentially useful patterns. It is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions. There are five steps in the KDD process.

i. **Data Selection**: This step focused on creating a target dataset, or a subset of variables or data samples, on which discovery is to be performed. At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

ii. **Data Pre-Processing**: This stage consists on the target data cleaning and pre-processing in order to obtain consistent data.

iii. **Data Transformation:** This stage of KDD process which focuses on transformation of data from one form to another so that data mining algorithms can be implemented easily. For this purpose, different data reduction and transformation methods are implemented on target data

iv. **Data mining:** It is an essential process where intelligent techniques are applied to extract potentially useful patterns.

v. **Interpretation/ Evaluation:** This stage focused on the interpretation and evaluation of the data mined patterns.

vi. **Knowledge representation:** Is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

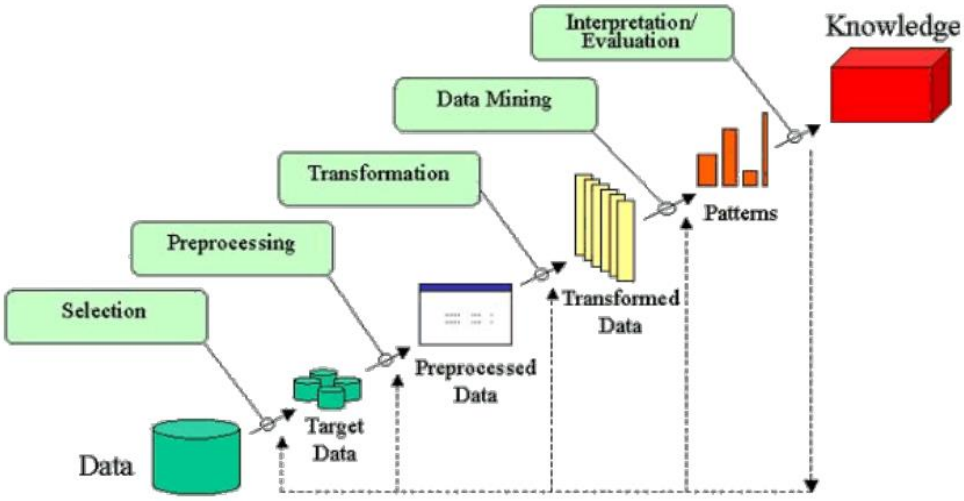

Figure2. 1: Knowledge Discovery Process

2.4.2. The CRISP-DM Process

CRISP-DM is the most used methodology for developing DM projects as referenced in (O. Marban, G. Mariscal, and J. Segovia, 2009). It is vendor-independent so which can be used with any DM tool to solve any DM problem. CRISPDM provides a uniform framework and guidelines for data miners. CRISP-DM also defines for each phase the tasks and the deliverables for each task. It consists of six phases or stages which are well structured and defined. These phases are displayed and described as follows.

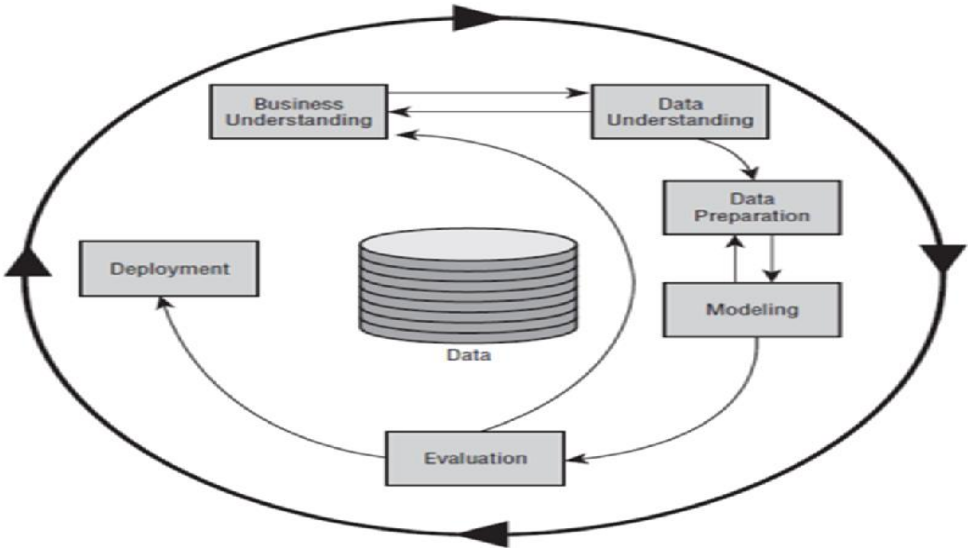

Figure2. 2: CRISP-DM process model.

i. **Business understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

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

iii. **Data preparation:** This phase covers all activities to construct the final dataset from the initial raw data. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

iv. **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

v. **Evaluation:** At this stage, before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

vi. **Deployment:** The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.4.3. Hybrid Model

.

It is developed by combining two process models KDD and CRISP-DM. In this process model six basic steps are passed to achieve the overall goals of data mining process. It includes understanding the problem domain, understanding the data, preparing of the data, Data mining, evaluating the discovered knowledge and finally use the discovered knowledge for real applications in the domain area. These phases are described as follows.

1. Understanding the problem domain.

The first phase involves working closely with domain experts to define the problem and determine the research goals, identifies key people, and learns about current solutions to the problem. It involves learning domain-specific terminology. A description of the problem including its restrictions is done. The research goals then need to be translated into the DM goals, and include initial selection of potential DM tools.

2. Understanding the data.

This phase includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the

DM goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc

3. Preparation of the data

This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data will be used as input for data mining tools of step 4, is decided. It may include sampling of data, running, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), and by derivation of new attributes (say by discretization), and by summarization of data (data granularization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.

4. Data mining

This is another key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. This step involves the use of several DM tools on data prepared in step First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.

5. Evaluation of the discovered knowledge.

This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the

impact of the discovered knowledge. Only the approved models are engaged. The entire DM process may be reconsidered to identify which alternative actions could have been taken to improve the results.

6. Using the discovered knowledge.

This step is completely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

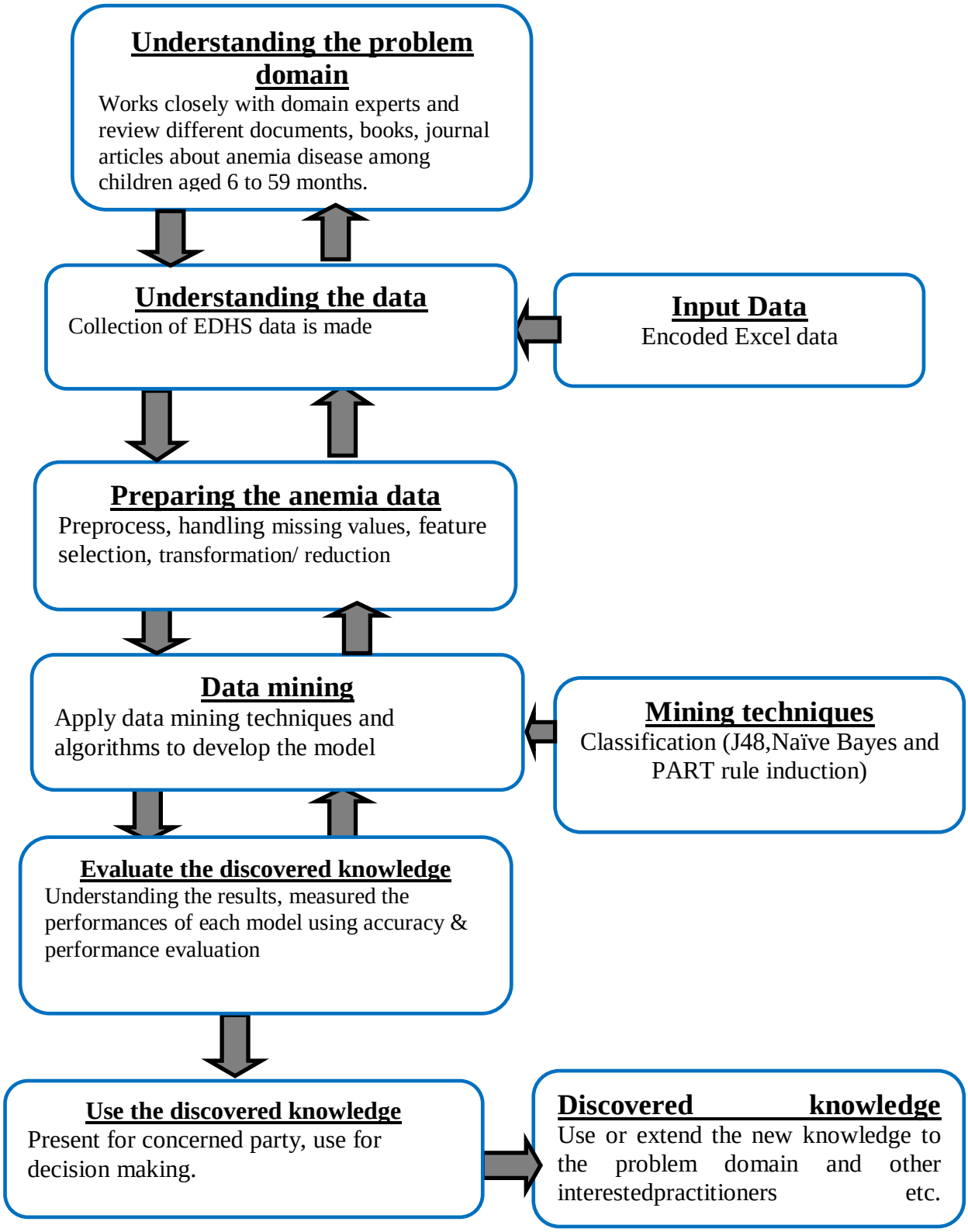

Figure2. 3 Stages of Hybrid Process Model (adopted Cios et al., 2007)

2.5. Review of Related Works

In this unit, some of the studies that have been conducted related to anemia disease using traditional statistical techniques are discussed.

One of the studies conducted on the prevalence of anemia was that of Bereket Geze et al (2018) entitled "Anemia and associated factors among children aged 6–23 months in Damot Sore District, Wolaita Zone, South Ethiopia". The objective of the study was to determine overall prevalence of anemia among children 6–23 months. A community-based cross-sectional study was carried out among 485 children of Damot Sore, South Ethiopia from March to April 2017. Data on socio-demographic, dietary, blood samples for hemoglobin level and malaria infection were collected. Socio-demographic and economic data were collected by interviewing mothers of the children during house to house data collection by using pretested and interviewer administered a questionnaire that was prepared in the English language, which later was translated into Wolayttattua-local language. Both descriptive and bivariate analyses were done and all variables having a *p*-value of 0.25 were selected for multivariable analyses. A multivariable logistic regression model was used to isolate independent predictors of anemia at a p-value less than 0.05. A principal component analysis was used to generate household wealth score, dietary diversity. Out of 522 sampled children, complete data were captured from 485 giving a response rate of 92.91%. For altitude and persons smoking in the house adjusted prevalence of anemia was 255(52.6%). The larger proportion, 128(26.4%) of children had moderate anemia. On multivariable logistic regression analyses, household food insecurity (AOR = 2.74(95% CI: 1.62–4.65)), poor dietary diversity (AOR = 2.86(95% CI: 1.73–4.7)), early or late initiation of complementary feeding (AOR = 2.0(95% CI: 1.23–3.60)), poor breastfeeding practice (AOR = 2.6(95% CI: 1.41–4.62)), and poor utilization of folic acid by mothers (AOR = 2.75(95% CI: 1.42–5.36)) were significantly associated with anemia. The researchers concluded that the Prevalence of anemia among children (6–23 months) was a severe public health problem in the study area. Most important predictors are suboptimal child feeding practices, household food insecurity,

and poor diet. Multi-sectorial efforts are needed to improve health and interventions targeting nutrition security are recommended.

Dereje Habte, (2013) conducted a study entitled "Maternal Risk Factors for Childhood Anemia in Ethiopia" the aim of the study was to identify the risk factors associated with childhood anemia in Ethiopia. The study population consisted of A total of 8260 children between the ages of 6-59 months. Data for this study was drawn from the third Ethiopian Demographic and Health Survey - 2011 (EDHS). To explore the predictors for the prevalence of anemia by using different possible related factors in mothers such as maternal age, maternal education level, maternal anemia, residence, and maternal wealth status. SPSS software (version 20.0) and Epi Info Version 3.5.3 were used in the analysis of the data.

The finding of the study discovered that Childhood anemia was shown to be a severe public health problem irrespective of urban residence, better wealth status or educational achievement. Despite the varying level of childhood anemia with some background characteristics like residence, wealth index and maternal education, all categories had more than 40% anemia prevalence. An anemia prevalence of 40% or more is defined as a severe public health problem. A Significant proportion of children had mild or moderate anemia while a few fell in to the severe anemia category.

Bentley, M and Griffiths, P, (2012) conducted a study the burden of anemia among women in India. They used the National Family Health Survey 1998/99 which provides nationally representative cross-sectional survey data on women's hemoglobin status, body weight, and diet, social, demographic and other household and individual level factors. The aim of the study was to explore the prevalence and determinants of anemia among women in Andhra Pradesh. The finding of the study discovered that prevalence of anemia was high among all women, and poor urban women had the highest rates and odds of being anemic.

Simbauranga et al., (2015) studded Prevalence and factors associated with severe anemia amongst under-five children hospitalized at Bugando Medical Centre, Mwanza, Tanzania. This study aimed to determine the prevalence and and morphological types of anaemia, as well as factors associated with severe anaemia in under-five children admitted at Bugando Medical Center (BMC).

The study used a hospital-based, cross-sectional study conducted between November 2012 and February 2013. Selected laboratory investigations were done on children admitted to BMC. Anemia was defined using WHO criteria. Results: A total of 448 under-five children were recruited into the study. The overall prevalence of anemia was 77.2 % (346/448) with mild, moderate and severe anemia being 16.5, 33 and 27.7 % respectively. Microcytic hypochromic anemia was detected in 37.5 % of the children with anemia. Of 239 children with moderate and severe anemia, 22.6 % (54/239) had iron deficiency anemia based on serum ferritin level less than12 g/ml. The factors associated with severe anemia included unemployment of the parent, malaria parasitaemia and presence of sickle hemoglobin.

The finding of the study were the prevalence of anemia among under-five children admitted at BMC was high. Iron deficiency anemia was the most common type. Factors associated with severe anemia were unemployment among caretakers, malaria parasitaemia and presence of sickle haemoglobin.

Getachew Mullu et al. (2017) conducted a study the Prevalence and determinants of anemia among pregnant women in Ethiopia. The study used Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline was followed for this systematic review and meta-analysis.. The meta-analysis was conducted using STATA 14 software. The pooled Meta logistic regression was computed to present the pooled prevalence and relative risks (RRs) of the determinate factors with 95% confidence interval (CI). Twenty studies were included in the meta-analysis with a total of 10, 281 pregnant women. The pooled prevalence of anemia among pregnant women in Ethiopia

was 31.66% (95% CI (26.20, 37.11)). Based on the pooled prevalence of the subgroup analysis result, the lowest prevalence of anemia among pregnant women was observed in Amhara region, 15.89% (95% CI (8.82, 22.96)) and the highest prevalence was in Somali region, 56.80% (95% CI (52.76, 60.84)). Primigravid (RR: 0.61 (95% CI: 0.53, 0.71)) and urban women (RR: 0.73 (95% CI: 0.60, 0.88)) were less likely to develop anemia. On the other hand, mothers with short pregnancy interval (RR: 2.14 (95% CI: 1.67, 2.74)) and malaria infection during pregnancy (RR: 1.94 (95% CI: 1.33, 2.82)) had higher risk to develop anemia. Almost one-third of pregnant women in Ethiopia were anemic. Statistically significant association was observed between anemia during pregnancy and residence, gravidity, pregnancy interval, and malaria infection during pregnancy. Regions with higher anemia prevalence among pregnant women should be given due emphasis. The concerned body should intervene on the identified factors to reduce the high prevalence of anemia among pregnant women.

Getachew Setotaw. (2013) studded predicting the status of anaemia in women aged 15-49 in Ethiopia. The aim of the research was to predict the status of anaemia in women aged 15-49 by applying data mining techniques using the 2011 EDHS dataset. The Knowledge Discovery in Database (KDD) which was applied to extract useful information from the dataset containing 6697 records.J48 Decision Tree and PART Rule induction were employed to build the predictive model. And also WEKA open software was used as a data mining tool to implement the experiments. The findings of the study revealed that all the models built using J48 decision tree and PART rule induction algorithms with all attributes have high classification accuracy and are generally comparable in predicting any anaemic cases. However, comparison that is based on the detailed performance measures suggests that the PART Unpruned model with all attributes perform better in predicting any anaemic cases with an accuracy of 95.2%.
Finally, women's age residence, Educlev, wealth, BMI, and Current contraceptive use were the most potential set of attributes.

2.6. Review of Research Papers/applications of data mining in health care

In this section, some of the researches that have been done by using data mining techniques are discussed.

One of the research conducted using data mining technique was that of (Abraham T, 2005) entitled "identify determinant risk factors of HIV infection and to find their association rules using data mining techniques": the case of Center for Disease Control and Prevention (CDC). To extract the unknown patterns among the variables under the study WEKA software was used. The researcher argued one of the important findings noticed under the study was a new insight about risk feeling of the clients and HIV test result. According to the researcher previously it was known that the clients whose reason for test is plan for future are associated with HIV-negative class. This truth has also been verified with experiment too. However, the experiment disclosed that people whose reason for test is having risk, suspect or symptoms is also associated to HIV-negative result with promising evidence. The researcher further noticed this was previously hidden information that domain experts were impressed to hear about. Accordingly, a client who has risky perception of oneself has a better chance to be uninfected.

A study which is done by Biset D. (2011), has applied data mining techniques for his master thesis to predict low birth weight on Ethiopian Demographic and Health Survey data sets. The aim of the study was to predict the status of low birth weight using EDHS 2005 (Ethiopia Demographic Health Survey) data set so as to build a model using data mining technique addressing the causes related with low birth weight. The researcher used CRISP-DM methodology. Two machine learning algorithms from WEKA software, J48 decision tree classifier and PART rule induction algorithms were selected for experiments.

The researcher compared the classification performance of the decision trees with tree pruning and without tree pruning, and found that tree pruning can significantly improve

decision tree's classification performance. In general, the results from the study as the researcher described were encouraging which can be used as decision support aid for health practitioners. He further indicated that the extracted rules in both the algorithms are very effective for the prediction of low birth weight. Lastly, from both algorithms the researcher observed that attributes such as, iodine contents in salt, antenatal visits during pregnancy, region, age of mother, mother 's educational level, marital status, wealth index, place of residence and numbers of birth order are the most influential factors to predict the status of low birth weight.

A research which conducted by Muluneh Endalew, (2011) to investigate the potential applicability of data mining techniques in exploring the prevalence of diarrheal disease using the data collected from the diarrheal disease control and training centre of African sub Region II in Tikur Anbessa Hospital. The researcher used the CRISP methodology, two machine learning algorithms from WEKA software such as J48 Decision Trees(DT) and Naïve Bayes(NB) classifiers, were implemented to classify diarrheal disease records on the basis of the values of attributes 'Treatment' and Type of Diarrhea'. Finally the experiments showed that J48 decision tree classifier has better classification and accuracy performance as compared to Naïve Bayes classifier. The researcher finalized that the study has showed that data mining techniques are important to support and scale up the effectiveness of health care services provision process.

A study conducted by (Senthilkumar D and Paulraj S ,2015) used data mining technology for Prediction of Low Birth Weight Infants and Its Risk Factors using the data collected from Indian health sector. Deeper understanding of the important factors highly associated with low birth weight, different data mining algorithms has been applied to the prediction of low birth weight child, including logistic regression, naïve Baye's, random forest, support vector machines (SVM), neural network and classification tree. The variables which were highly influenced in the predication of low birth weight are Mother's last weight (pounds) before becoming pregnant, Mothers age, Number of

physician visits during the first trimester, Number of previous premature labors. Classification tree performed best compared with other algorithms.

Another research which done by Behailu Gebre Mariam and Tesfahun Haile Mariam,(2015) applied data mining techniques for Predicting CD4 Status of Patients on ART in Jimma and Bonga Hospitals, Ethiopia. The study followed the CRISP-DM data mining methodology which has six phases: business understanding, data understanding, data preparation, model building, evaluation and deployment. For this study, data was taken from two hospitals of the south west of Ethiopia; Jimma and Bonga hospitals. Classification algorithm was used to predict CD4 status of the patients those who are following ART therapy. J48 is a technique used for building classification and PART is used to compare the result of J48 algorithm.

Classification done using J48 decision tree is the best model as compared to PART rule algorithm and that can be used for prediction. From the model built it is possible to conclude that attributes like: Eligible reason, ART status, ART start year, OA weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status, Past ARV are the most determining factors of CD4 status.

The ten top determinant attributes are Eligible reason, ART status, ART start year, OA(Osteoarthritis) weight, OAWHO stage, Current regimen, Family planning, Functional status, Marital status and Past ARV consecutively. From the attributes ranked for predicting CD4 status of patients who are following ART, Eligible reason attribute is became the most determinant attribute whereas educational level became the least determinant attribute.

A study which is conducted by (Hung Y et.al, 2011) has applied data mining techniques (C4.5, B1 and naïve byes) to a diabetic patient database. The purpose of the study was to identify determinant factors influencing diabetes control; Predicting individuals in the population with poor diabetes control status based on physiological and examination

factors with the purpose was to advance the quality of treatment by automating the handling of routine situations, particularly blood glucose control, and ensuring a better quality of life by providing support from expert. Data integration has done to merge separated sources available and data transformation is done in order to make the techniques work on discrete variables. Identification and removal of irrelevant and redundant information in other word feature selection is done to improve the efficiency of the data mining techniques.

A study conducted by (Shegaw Anagaw, 2002) applied data mining technology to predict the risk of child mortality based up on community-based epidemiological datasets gathered by the BRHP epidemiological study. The methodology used for this study had three basic steps, these are collecting of data, data preparation and model building and testing. The necessary data was Selected and extracted from the ten years observation dataset of the BRHP epidemiological study. Models were built and tested by using a sample dataset of 1100 records of both alive and Died children.

Several decision tree models and neural network were built and tested for their classification accuracy and many models with encouraging results were obtained. The two data mining methods used in this research work have proved to yield comparably sufficient results for practical use as far as misclassification rates come into consideration. However, unlike the neural network models, the results obtained by using the decision tree approach provided simple rules that can be used by non-technical health care professionals to identify cases for which the rule is applicable.

Several classifiers were also constructed by using See5 decision tree software. From those classifiers, the best classifier was achieved when the ruleset and adaptive boosting options were used. The classifier was built by using the following attributes: “ENVIRN”, “AGE”, “SEX”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”,” HHMEMBAVE”, “HHLIVESTOK”, AND “WINDOWS”. This classifier resulted with an accuracy of 95% (i.e. it classified 942 of

the 995 training cases correct) on training cases and it achieved 95% accuracy (classified 105 of the 111 test cases correct) on test cases.

The results obtained in this research work have proved the potential applicability of data mining technology to predict child mortality patterns based solely on demographic, parental, environmental, and epidemiological factors. The encouraging results obtained from both neural networks and decision trees indicate that data mining is really a technology that should be considered to support child health care prevention and control activities at the district of Butajira in particular, and at a national level in general.

Apart from the above researches as to the knowledge of the researcher no study was done by applying data mining techniques to predict the status of anemia among children aged 6 to 59 months in Ethiopia using EDHS data set. Hence it is the aim of this research to apply data mining techniques in order to predict the states of anemia among children aged 6 to 59 months.

CHAPTER THREE

3. RESEARCH METHODOLOGY

The purpose to specify methodology is to provide insight with fundamental knowledge of data modeling and design; the tools and techniques of data analysis using data mining technology; to inform beneficiaries with data mining concepts, and techniques; and to prepare data for further analysis in the database.

3.1. Study area

In this study, the researcher was mainly focused on predicting the status of anemia among children aged 6 to 59 months on EDHS 2016 datasets. The 2016 EDHS survey covers both urban and rural areas of Ethiopia. A total of 11 geographic/administrative regions, that is, nine regional states and two city administrations, namely: Afar, Amhara, Gambella, Harrari, Oromia, Somali, Benshangul - Gumuz, Southern Nations, Nationalities and Peoples (SNNP), Tigray, and Addis Ababa, and Dire Dawa were included in the study (EDHS, 2016).

3.2. Data collection

This research was generally based on secondary data extracted from the Ethiopia demographic and Health Survey (EDHS) of 2016, the most current national dataset on anemia testing that is accessible (as of July 2017). This was conducted by the Central Statistical Agency (CSA) under the worldwide MEASURE DHS project. The researcher obtained such EDHS 2016 datasets from the Ethiopian Central Statistical Agency (CSA) after submitting a formal letter written by the School of computing.

In this research, three files such as Household, Individual, and Couples ‘stored in the 2016 EDHS dataset were used as a source of the data. All these files were available in

SPSS format contained a total of 41,392 records (rows) and 1245 attributes (columns). Therefore, no further data collection mechanisms are evolved as the collected data are ample enough to undertake the planned research.

3.3. Research design

The study employed a hybrid data mining process model approach to build predictive models using data mining techniques by applying a set of classifier algorithms on the EDHS 2016 datasets. In this study, hybrid approach is built by combining knowledge discovery in databases (KDD) and cross industry standard process for data mining (CRISP-DM) process models. This method was selected for this specific study because of various reasons.

- It provides more general, research-oriented description of the steps.
- It underlines on the consistent aspects of the process, drawing experiences from the previous models, and
- It supports both industrial and academia data mining project standards.

In the selected data mining approach six main activities are considered for the development of the required model developed by Cios et al. (2007) is considered because this model combines both academics and industry aspects. These include, understanding the problem domain, understanding the data, preparing the data for mining, mining the prepared data, evaluating or testing the discovered knowledge on new datasets and using or applying the discovered knowledge for decision making.

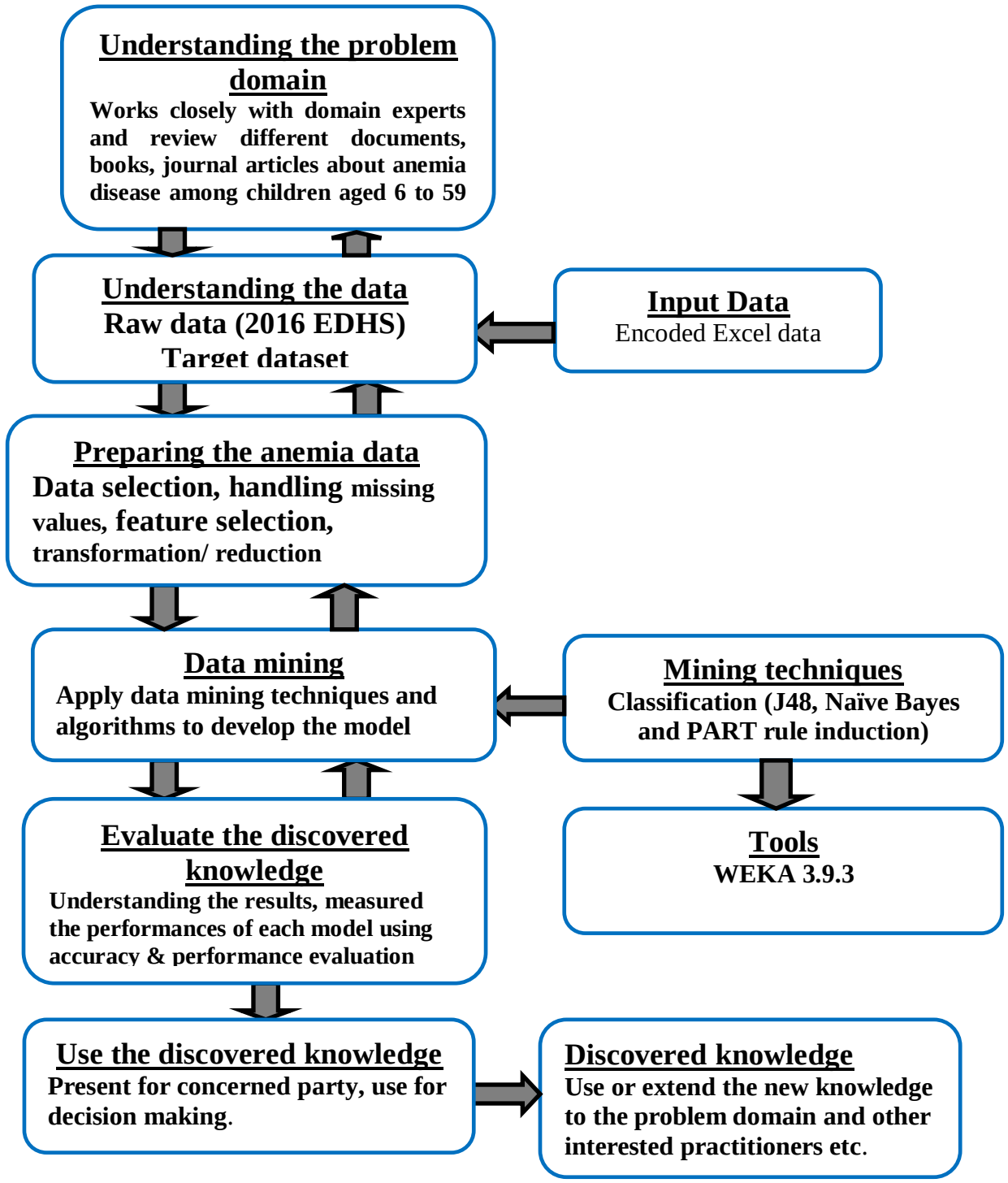

Figure3. 1: Research design

3.3.1. Understanding the problem domain

This initial phase focuses on understanding the project objectives and requirements from a business point of view, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives (Cios et al., 2007). Hence the researcher was communicating and works closely with domain experts and review different documents, books, journal articles and conference papers that focus on data mining techniques and applications in the healthcare to define the problem as depicted in Section 1.2 of chapter one and an effort is made to determine the project goals, and oversee current solutions to the problem. Then, the project goals were converted into data mining goals. Serious of interviews with the organization's domain experts has enabled the researcher to define the data mining problem. This has led the researcher to conduct a research by applying data mining to predict the status of anemia among children aged 6 to 59 months. For this reason, classification technique was applied to develop a model.

3.3.2. Data understanding

In any data mining task, the first step is clear understanding of the problem to be solved and this has been already addressed in Section 1.2 of chapter one of this study which helped to know what data is required to perform the task. Domain experts were consulted to have a clear understanding of the data. The most important thing in any data mining project is the data itself and the source of the data.

According to (Cios et al., 2007) the data understanding phase primarily focuses on creating a target dataset with selected sets of variables that is related to discovery process. Without understanding the existing data, it is difficult to describe the target dataset from the source since the world data is unclean and not suitable at the source to run mining process.

3.3.2.1. Source and Description of the Data

As described in section 3.1 of chapter three, the main data source for the purpose of this thesis work was 2016 Ethiopia Demographic and Health Survey (2016 EDHS) implemented by the Central Statistical Agency (CSA). The primary objective of the 2016 EDHS project is to provide up-to-date estimates of key demographic and health indicators.

A total of 8603 records having 21 attributes were employed for this thesis work. The agency records background information and health data in SPSS format. Hence, the researcher had converted the data in an excel format then preprocessing techniques were applied to make it appropriate for mining process. Accordingly, the original attributes with their description and data type are presented in table 3.7.

3.3.2.2. Data Quality Assurance

In general, the researcher was satisfied with the reliability of data and completeness of the records. However, the collected data contains missing, repeated and irrelevant data. The original dataset also contained large number of attributes, but some of these are repeated information that could be found in other attributes. According to (Han J and Kamber, M. 2006) data quality can be measured in terms of accuracy, completeness, consistency, timeliness, believability and interpretability. Since data quality is the reliability and effectiveness of data, particularly in a data warehouse.

3.3.3. Data Preparation

As stated in Two Crows Corporation, (2005) one of the most important tasks in data mining is preparing the data in a way that is appropriate for the specific data mining tool or software package to be used. Data preparation includes data selection, data cleaning, data construction, data integration and dataset formatting.

In this stage, the final dataset was prepared from the initial raw data. Accordingly, missing values, outliers and noisy data are identified and handled, and data transformation/ reduction activities are also undertaken in this stage. The researcher used WEKA 3.9.3 and MS Excel for data preparation and analysis task because it has the capability of filtering attribute with different values. Finally, the collected data were arranged and converted into a form CSV and ARFF file that is suitable for the data-mining tool selected. Due to the nature of the datasets classification data mining technique was selected to classify the sample dataset using algorithms such as J48 Decision Tree, Naïve Bayes and PART Rule Induction with data mining tools of WEKA.

According to Han J. and Kamber M. (2006) recommended at attention should not be ignored to clean data for knowledge mining because the real-world data is highly vulnerable to noisy, inconsistency and incompleteness. The researchers state that the necessitate for data preparation is, today's real-world data are greatly exposed to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, various sources. Low quality data will lead to low quality mining results.

This phase includes a number of steps to provide the final dataset for modeling. As shown in Figure 3.2. It contains data selection, cleaning, construction, integration and formatting.

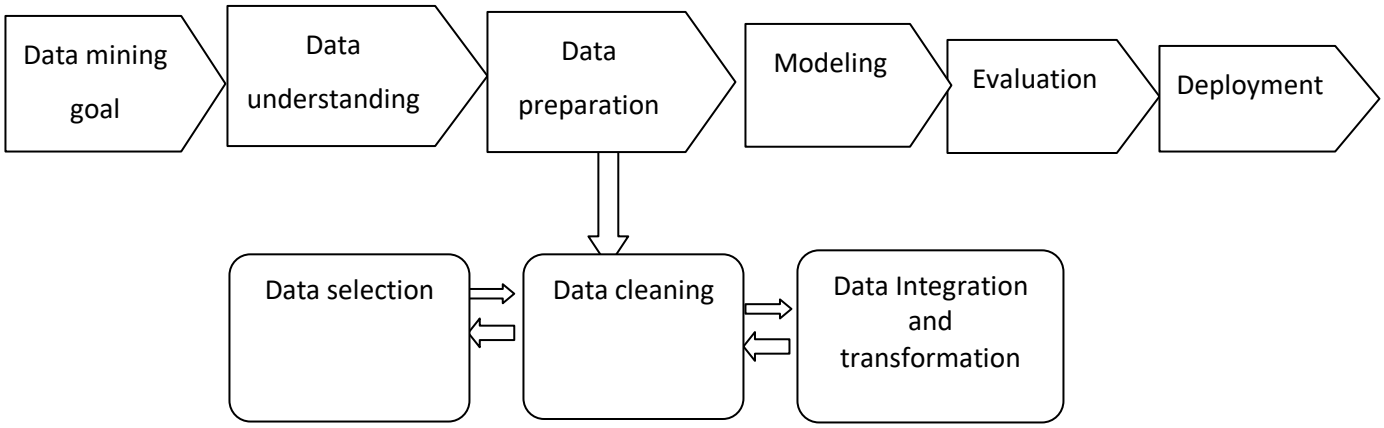

Figure3. 2: Data Preparation phase

3.3.3.1. Data selection

Data selection is the process of selecting on the right data from the database on which the tools in data mining can be used to extract useful information, knowledge and pattern from the provided raw data.

As it has been described under the objective of the study on the first chapter, the aim of this study is to construct anemia predictive model using data mining techniques so as to identify the influential factors of anemia in children aged 6-59 months using the 2016 Ethiopia Demographic and Health Survey (EDHS) survey datasets collected by CSA. And the source data employed for this research was those three files of the 2016 EDHS datasets indicated in data collection section on chapter one. The 2016 EDHS survey dataset that contains women aged 15-49, children aged 6-59 months, and men aged 15-59 was originally stored in SPSS format. In the source dataset a total of 1245 attributes (columns) and 41,392 records (rows) were identified.

After collecting the SPSS format of the original survey dataset, Microsoft Excel is used for selecting the target data set required for this study. After the elimination of irrelevant and unnecessary data we had 8603 data set ready for this study. In this first step, from the source dataset, data only children aged 6-59 months was selected to achieve the objectives of this study.

3.3.3.2. Attribute selection

As it is stated in section 3.3.3.1, the original dataset from which the target dataset for this research work has been chosen consisted of a total of 1245 attributes. But all these attributes found in the original dataset were not needed for this study. Therefore, only related attributes to predict status of child anemia were considered for the specific learning task, we have excluded many features that are not necessary for this study.

Large number of attributes or features is not useful because they are either irrelevant or redundant to the prediction (Kim, Y. Street, and Menczer, F, 2000). Whereas there are different methods that can be applied to attribute selection, the researcher used the manual way of selecting attributes only data's relevant and believed by domain experts to the analysis task are selected. Domain experts have knowledge from public health field specifically nutritionist whose list is shown in Appendex1, and wide review of all attributes detailed in 2016 EDHS documentation files, 20 potential independent attributes are selected based on their background characteristics which could determine the status of anemia in children aged 6-59 months.

These potential attributes are grouped under Socio-demographic attributes of maternal, household and child attributes (11 attributes), nutritional status of both maternal and child attributes (6 attributes), and health related (3 attributes), The researcher more analyzed each of these grouped attributes and introduces the selected set of attributes in each category that are lastly used in the experiment phase.

- **Socio-demographic attributes of maternal, household and children aged 6-59 months.**

The following table shows Socio-demographic attributes of maternal, household and children aged 6-59 months with its description

Table 3.1: Socio demographic attribute

No	Name of attribute	Description
1	MAGE	Maternal age
2	REGION	Region where the survey conducted
3	HHNUM	Number of household's members
4	RESIDENCE	Type of place of residence

5	MEDUCLEV	Maternal education level
6	WEALTH	Wealth status of household
7	FEDUCLEV	Father's education level
8	MOCCUP	Maternal occupation
9	SEX	Sex of child
10	CHILD AGE	Child's age in months
11	BIRTH INTERVAL	Child's birth interval

All these attributes are categorical (either flag, ordinal or nominal), except AGE attributes of both maternal and child which is continuous (scale) variable.

AGE represents the age data of maternal/mother and child tested for anemia. In the 2016 EDHS, it has been top-coded to 49 years and 59 months for maternal and child in original data and ranges from 15 to 49 years and 6 to 59 months respectively. The age of maternal is classified by five-year age groups. This attribute is categorized into five groups: 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, whereas the age of child attribute is categorized into two groups: 6-24months(<=2years) and 25-59months (>2years).

REGION: The region attribute indicates the location of mothers and child. This attribute contains a total of 11 administrative region of the country; such as Tigray, Afar, Amhara, Oromyia, Somali, Benshangul-Gumuz, SNNP, Gambella, Harari, Addis Ababa, and Dire Dawa.

HHNUM: represents number of household, either number of households 2-5(between 2 and 5), number of households greater than or equals to 6).

RESIDENCE: represents the type of place of respondents' residence, either urban or rural.

MEDUCLEV and FEDUCLEV: This attribute tells the level of education both mother/maternal and father's education. Mother's and father's education are indirectly

related to a child's health. It is nominal attribute that contains four values (No education, Primary, Secondary, and Higher)

WEALTH is the categorical variable that shows the wealth status of the respondents. It accepts values of 1 to 3 for poor, middle, rich respectively.

EMPLOY represents the current employment status of the child's mother. It is a flag attribute accepting 0 (No) or 1 (Yes) values.

MOTHOCCU represents mother's occupation is an indicator of economic level in relation to hemoglobin level of a child with distinct values of attribute is not working and working.

SEX indicates the gender of each child either male or female, with a value equal to Male or Female.

BIRTHINTERVAL represents child's birth interval, either birth interval <36months, or >=36months).

- **Nutritional status attributes of both maternal and child aged 6 to 59 months**

The following table shows nutritional status attributes of maternal, and child under five age with its description.

Table3.2: Nutritional status attributes

No	Name of attribute	Description
1	MBMI	Maternal Body Mass Index
2	MANEM	Maternal hemoglobin level
3	WHZ	Nutrition status of child or Weight/Height standard deviation (new WHO)
4	BREASTF	Currently breastfeeding status

5	CHILD SIZE	Size of child at birth

All these attributes are categorical (either flag, ordinal or nominal), except BMI and CHILDSIZE attributes of maternal and child respectively which is continuous (scale) variable.

CHILDSIZE: Size of a child at birth. The distinct values of this attributes are large (>4 kg), normal (2.5-4 kg) and small (<2.5 kg).

MBI: Maternal body mass index ranges from low/under weight (<18.5), normal weight (18.5 – 24.9), and overweight (>=25.0). BMI is a continuous (scale) variable. BMI is a measure of body mass index and person‘s height and weight (CSA [Ethiopia], 2011). Based on insights taken from literature review in which the nutritional status is considered as an important determinant factor for the prevalence of anemia.
WHZ attribute: The WHZ index measures body mass in relation to body height or length; it describes current nutritional status of child. Final WHZ attribute discredited categories are <-2SD (Wasted), -2SD to 2SD (Normal) and >2SD (Over)

MANEM: This attribute is an indicator for anemia level of a maternal/ mother and an indicator for anemia status of a child. The distinct values of this attributes are anemic (<12g/dl) and not anemic(>12g/dl).

BREASTF: Currently breastfeeding status of child, it is a flag attribute accepting No or Yes values.

- **Health related attributes**

The following table shows health related attributes of maternal, and children aged 6-59 months with its description.

Table3. 3: Health related attributes

No	Name of attribute	Description
1	NUMANV	Antenatal visits during pregnancy
2	Had diarrhea recently	Children Had diarrhea recently
3	VITAMINA1	Received Vitamin A1 (most recent)
4	PNC	postnatal care of a child

ANV: this attribute represents the status of maternal antenatal visits during pregnancy. It contains two distinct values no antenatal visit, <4, and >=4 antenatal visit using WHO standards.

PNC: this attribute is representing the status of health after birth. The distinct values of this attributes are <3 and >=3.

DIARRHEA: is an attribute used to check that are the current status of child is infected or not. The distinct values of this attribute are Yes and No

VITAMINA1: child's Received Vitamin A1 (most recent), it is a flag attribute accepting No or Yes values.

3.3.3.3. Data cleaning

Data cleaning helps to clean the data by filling in the missing values removes noise and corrects data inconsistency. Usually, real world database contains incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006). Incomplete data can occur for a number of reasons having incorrect attribute values, data collection instruments used may be faulty which may result in noisy data, human or computer errors occurring at data entry, and missing the most important attribute values. Consequently, data cleaning has become a must in order

to improve the quality of data so as to improve the performance of the accuracy and efficiency of the data mining techniques.

As a result, the researcher use MS-Excel 2010 built-in functions like search and replace, filtering, and auto fill mechanisms, and WEKA to identify and fill missing value.

i. Handling of Missing Values

Missing values refer to the values for one or more attributes in a data that do not exist. In real world application data are rarely complete. Missing attributes values in the dataset is most likely related with unavailability of interesting information, lack of knowledge on the importance of data at the time of data entry, misunderstanding of the data, the respondents him/herself may reject to answer certain questions or they may not know the answer accurately or may answer in an unexpected way (Kantardzic M, 2003).

As it is recommended in (Han J and Kamber M. 2006) Missing values for an attribute could be filled using different method. Among these; removing the tuple with missing value, fill in the missing value manually, use a global constant to fill in the missing value, use the attribute mean for all samples belonging to the same class as the given tuple, or use the most probable value to fill in the missing value. From the available attributes "maternal anemia level", "maternal BMI" , "Waiting time for birth", "Child age", "Child size", "Had diarrhea recently", and "Vitamin A1" were found missing value.

As it is stated in section 3.2 of this study, three files of the 2016 EDHS dataset such as Household, Individual, and Couples' which covered a total of 41, 394 records and 1245 attributes were employed as a source data. However, while looking each individual file, some attributes values which are existed in one file missed in another files. For instance, in the Household file, vitamin A1 had 45% had missing values, to control such difficulties of missing values, the researcher used the data imputation method to substitute the missing values rather than just rejecting those attributes which have highest missing values.

After using such techniques of substituting missing values mainly for those attributes which had highest missing values, therefore, in the final dataset of this research work, 10 out of 20 attributes have missing values. Table 3.4 shows the value labels for missing values in each related field.

Table3.4: Missing values in each related field.

No	Attribute Name	No of Missing Value	% of Missing value
1	MBI	314	4
2	MOTHANEM	448	5
3	FEDULEV	562	6.5
4	BIRTHINTERVAL	1210	14
5	NUMANV	1118	13
6	CHILDSIZE	79	0.92
7	DIARRHEA	13	0.15
8	VITAMINA1	3869	45
9	WHZ	612	7
10	CHILD AGE	808	9

To solve the problem of attributes with missing values in the dataset formed for a specified data mining task, since all attributes shown in Table 3.4 have below 50% missing values, the researcher used modal and mean value approach for substituting the missing after identifying their attribute type.

Based on the method, in the final dataset of this study, since all attributes except age attribute with missing values are nominal categorical variables, for any missing value in such field, the modal (most frequent) value was substituted. The modal values for the above nominal attributes are shown in Table 3.5.

Table3.5: Modal values representing missing values

No	Attribute Name	Modal Value
1	MBI	Normal
2	MANEM	Not anemic
3	FEDULEV	No education
4	CHILDAGE	<=2YEARS
5	BIRTHINTERVAL	<36months
6	NUMANV	Number of antenatal visit
7	CHILDSIZE	Normal
8	Had diarrhea recently	No
9	VITAMINA1	No
10	WHZ	Normal

ii. Handling Outlier Value

The data stored in a database may reflect outlier – noise, exceptional case, or incomplete data object and random error in a measure of variable. These incorrect attribute values may be due to data encoding problems, wrong data collection, and irregularity in naming convention or technology limitation (Han J and Kamber, M. 2006). The authors have also explained four basic methods for handling of noise data. These are binning method, clustering, regression and combined computer and human inspection.

In this study, the researcher has identified and detected noise or outlier value from the anemia data. With the support of domain experts, the identified outlier was corrected manually. Therefore, a combined effort of the researcher and domain expert were taken to identify and right the problem of the inadequate, noise or outlier.

3.3.3.4. Data transformation

According to Han J and Kamber M. (2006) in data transformation; data are transformed or consolidated into forms appropriate for mining process. It involves smoothing, aggregation, generalization, normalization, discretization, and attributes construction. To make the dataset appropriate for this study data discretization technique is applied on numeric attributes to minimize distinct values of attributes, dimensionality reduction is

also used to reduce the size of the dataset and attribute selection method is the last method applied to remove weakly relevant attributes.

3.3.3.4.1. Data Discretization

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute in to intervals (Han J and Kamber, M. 2006) Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels there by reduces and simplifies the original data. This leads to a brief, easy to use, knowledge-level representation of mining results.

Discretization is the process of converting continuous valued variables to discrete values where limited numbers of labels are used to represent the original variables. The discrete values can have a limited number of intervals in a continuous spectrum, whereas continuous values can be infinitely many (Daniel T. L, 2005). In this research, household number, postnatal care, birth interval, child age, number of antenatal visits, Child size, maternal anemia level, WHZ and WEALTH attribute is re-binned into a new nominal attribute.

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

Discrediting the values of size of child at birth attribute: The values are very large and larger than average is more than 4 kg, average (normal) is equal to 2.5-4 kg and smaller than average and very small is less than 2.5 kg. This is done by defining a portion of the values through explicit data grouping as presented in (Table 3.6). It shows that discretization of size of a child at birth. Except second one, Normal, the first two values

very large and larger than average are merged into large and last two values smaller than average and very small are merged into small.

Discrediting the values of wealth attribute: The values are poorest, poorer, middle, richer and richest is re-binned into a new nominal attribute with values of poor for poorest and poorer, middle for middle and rich for richer and richest. It reduces the number of categories for wealth attribute from five to three and makes interpretation much easier.

The distinct values of household member attributes are 2-5, and >=6 and the distinct values of child age attributes 0,1,2,3,4 and 5 are the numeric values (such as age and household member) are converted to nominal, such as 2-5 and 6 or more for household member, <=2 years and >2 years for child age.

Maternal age: The age attribute was available in two formats; discrete (grouped age) and continuous value. Therefore, the grouped age attribute is selected.

Discrediting the values of MBI attribute: Mother's MBI is an indicator for nutritional status of a mother. The mother BMI discretized categories are <18.5 (underweight), 18.5-24.9 (normal) and >=25 (over weight). The rest discretized attributes are discussed on the following table 3.6.

Table3.6: Data Discretization

Attribute name	Old value	New/replaced value
Size of child	Very Large	Large
	Larger than Average	
	Normal	Normal
	Smaller than Average	Small
	Very Small	
WEALTH	Poorest	Poor
	Poorer	
	Middle	Middle
	Richer	Rich
	Richest	
MBI	<18.5	Under weight
	18.5 – 24.9	Normal
	>=25	Over weight
WHZ	<-2SD	Wasted
	-2SD – 2SD	Normal
	>2SD	Over
HOSEHOLDNUM	2,3,4,5,6,7,8,9,10…	2 -5
		6 or more
CHILDAGE	0,1,2,3,4,5	<=2years
		>2years
MOTHANEM	Not anemic	Not anemic
	Mild	Anemic
	Moderate	
	Severe	
ANV	Visit 0,1,2,3	<4
	Visit 4,5,6,..	>=4

Birth interval	10, 11,12,13…36 months	<36 months
	36,37,38…..months	>=36 months
PNC	1,2,3,4,5	<3, >=3

Table3.7: Final list of selected attributes with their descriptions

No	Attribute	Data type	Description
1	MAGE	Ordinal	Maternal age
2	REGION	Nominal	Area where the survey conducted
3	HOUSEHNUM	Ordinal	Number of households
4	RESIDENCE	Nominal	Type of place of residence
5	MATEDUC	Nominal	Maternal education level
6	WEALTH	Nominal	Wealth status of household
7	FEDUCLEV	Nominal	Husband education level
8	MOCCUP	Nominal	Maternal occupation
9	SEX	Nominal	Sex of child
10	CHILDAGE	Nominal	Child's age in months
11	BIRTHINTERVAL	Nominal	Preceding Child's birth interval
12	MBMI	Ordinal	Maternal Body Mass Index
13	MANEM	Nominal	Maternal hemoglobin level
14	WHZ	Nominal	Weight/Height standard deviation (new WHO)
15	VITAMINA1	Flag	Received Vitamin A1 (most recent)
16	BREASTF	Flag	Currently breastfeeding status
17	CHILDSIZE	Ordinal	Size of child at birth
18	ANV	Ordinal	Antenatal visits during pregnancy
19	DIARRHEA	Flag	Had diarrhea recently
20	PNC	Nominal	Child postnatal care

3.3.3.4.2.Target/Class attributes

Some data mining techniques need predefined classes in order to train and build classification models. In such cases, the data preparation task is unfinished until the target attributes are well-defined. In other words, the training set should be pre-classified so that the data mining algorithms know what the user is looking for. With respect to target attribute, as the purpose of this research was to predict the status of anemia among children aged 6 to 59 months, the target attribute selected in this study is anemia level. This attribute primarily has four different values such as not anemic, mild, moderate, and severe. However, to make the interpretation of the final result much easier, like other independent variables, the original anemia level is re-binned into a new structure of nominal attribute with values of any anemic (mild, moderate and severe) and not anemic. These new nominal attributes have been changed by using the same definitions of the 2016 EDHS documentation files that are similar with the WHO definitions for the rate of anemia in the world.

In general, the target attribute used in this research work is anemia level that has two values, namely, anemic and not anemic. This attribute is considered as dependent variable whereas the rest of the attributes specified in Table 3.7 are the independent attributes for this research work.

3.3.4. Data mining

This step involves usage of the planned data mining techniques, tools and selection of the new ones. Data mining tools include many types of algorithms, preprocessing techniques, and data mining elements. In the study, a classification technique is used to develop the model capable of answering the stated problems. The training and testing procedures are designed and the data model is constructed using the chosen data mining tools. The researcher used J48 decision tree, Naïve Bayes and PART Rule Induction classification

algorithms, data preprocessing and model development tools such as, MS excel, WEKA 3.9.3 and also documentation tools such as MS-Office packages.

This research is conducted by adopting three data mining classification algorithms, which includes Decision Tree (DT) with J48 classifier to generate rules sets, PART rule induction and Naïve Bayes algorithm to predict class membership probabilities were implemented. Classification is one of the major data mining tasks. So, this task is accomplished by generating a predictive model of data and interpreting the model regularly to provide information for selective labeled classes in data. Each algorithm used in this study is described in the following subsections. These algorithms were selected due to their popularity and usefulness in solving data mining classification problems. Moreover, their advantages such as easy to interpret/ understand and visualize the result, fast at classifying unknown records/new instances, and easily handles all types of attributes and missing values, implement and use and numeric attributes compared to other classification techniques are taken as other reason for selecting these classification techniques.

3.3.4.1. J48 decision tree algorithm

The decision tree algorithm used in this research is J48 algorithm, which is one of the most common decision tree construction algorithms (Witten and Frank, 2005) which is the successor of ID3 (Iterative Dichotomiser,) C4.5. J48 Decision tree is a popular utility that involves decision-based classification and adaptive learning over a training set. Whitten and Frank.(2000) further stated J48 algorithm of decision tree technique is one of classification and prediction algorithms which support both numeric and nominal predicators and nominal class attribute values.

3.3.4.2. Naïve Bayes algorithm

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class (Ng

AYJ, M. I., 2002). Naive Bayes is a type of supervised-learning module that contains examples of the input-target mapping the model tries to learn. Such models make predictions about new data based on the examination of previous data. The Naive Bayes algorithm uses the mathematics of Bayes' Theorem to make its predictions (Ng AYJ, M. I., 2002).

3.3.4.3. PART Rule induction

The rule induction classifier used in this study is PART, one of the most commonly rule based classifications method. Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form IF condition THEN conclusion (Han J, Kamber, M., 2012). PART is usually called a separate-and-conquer technique because it identifies a rule that covers instances in the class (and excludes ones not in the class), separates them out, and continues on those that are left. Such algorithms have been used as the basis of many systems that generate rules. The algorithm generates sets of rules called ‘decision lists’ which are ordered set of rules. PART obtains rules from partial decision trees using J48, and builds a partial C4.5 decision tree and converts the "best" leaf into a rule (Witten and Frank, 2005).

3.3.4.4. Validation techniques (Test Options)

All the identified algorithms experiment was tested with the option of using 10-fold cross-validation (the classifier evaluated using the number of folds that are entered in the folds text field). The default 10-fold cross validation was chosen and used for building the model and test the performance of the model. The researcher decided to use 10- fold test option due to its ability to perform widespread tests on many datasets with different learning techniques and 10 is the right number of folds to get the best estimate of error and there is also some theoretical evidence that backs this up.

In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. Each time it is trained on nine folds and tested on the remaining single fold. This validation technique was executed to minimize the bias associated with the random sampling of the training and testing data samples through repeating the experiments ten times.

3.3.4.5. Model Evaluation

This is the most significant step in any data mining task and the key to making real progress in data mining, because it gives us a clear view regarding the strength of each model individually, and allows us to compare different experimentation models based on common performance measures (Weiss, Sholom M. and Zhang, Tong, 2003). Accuracy is the basic performance measure, which computes, the percent of correctly classified instances in the test set.

The other performance measurement used to compare classification algorithm is confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The standard performance measures of confusion matrix that were used for evaluation such as correctness accuracy, TP Rate(sensitivity), TN Rate(specificity), precision, recall, F-measure, and ROC (Receiver Operating Characteristic). Each of performance measures were used where appropriate in the analysis of the performances.

In each experiment confusion matrix was applied to check how much data was correctly classified in true positive (TP) correctly classified as positive cases and false negative (FN) child with anemia falsely classified as not anemic in the study and also false positive (FP) healthy/ not anemic child falsely classified in child with anemic case and true negative (TN) correctly classified as negative case. Table 3.1 shows the two dimensional confusion matrix.

Table 3.8: Two-dimensional confusion matrixes.

	Predicted class		
Actual class		Yes	No
	Yes	TP	FN
	No	FP	TN

To calculate the above performance measures such as accuracy, TP, FN, FP, TN, Precision, Recall, F-measure and ROC area the confusion matrix results are applied. For example, the following are the general formulas of each performance measures.

Accuracy: To gain the accuracy of a classifier is by dividing the total correctly classified positives and negatives instance by the total number of samples. These measures are defined as

Accuracy = $\frac{TP+TN}{TP+FN+FP+TN}$

TP (sensitivity): Sensitivity is also referred to as the true positive rate is defined as its ability to correctly identify actual cases, i.e. for this study. To classify the positive cases correctly is sensitivity.

Sensitivity =$\frac{TP}{TP+FN}$

TN (specificity): Specificity is the true negative rate is defined as ability to correctly identify negative cases which are correctly identified. These measures can be computed as: Specificity = $\frac{TN}{TN+FP}$

Precision: Can be thought of as a measure of exactness (i.e, what percentage of tuples as positive is actually positive.

Precision= Positive Predictive value= $\frac{TP}{TP+FP}$

Recall: Is a measure of completeness (what percentage of positive tuples is labeled as positive. It is the same as sensitivity (or true positive rate). These measures are defined as

Recall = Sensitivity= TP Rate= $\frac{TP}{TP+FN}$

F-Measure: it is the inverse relationship between precision and recall, and calculated as harmonic mean between precision and recall.

F-measure= $\frac{2(Precision*Recall}{Precision+Recall}$

ROC curve: ROC curve allows us to visualize the trade-off between the true-positive (TP) rate and the false-positive (FP) rate at which the model can accurately recognize positive cases versus the rate at which it mistakenly identifies negative cases as positive for different portions of the test set. ROC curve for a given model shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) (Jiawei Han, 2001). It is performed by drawing curve in two dimensional spaces by representing vertical axis and horizontal axis, vertical axis for true-positive rate and the horizontal axis for false-positive rate. To assess the accuracy of a model, one can measure the area under the curve which is a portion of the area of the unit square and its value is ranged from 0-1.

The model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger evidence for a positive actual state (1.00) (Cios Krzysztof J, 2007).

3.3.5. Evaluation of the discovered knowledge

This step includes understanding the results, checking whether the discovered knowledge is novel and interesting, impact of the discovered knowledge and interpretation of the results by domain experts. Performance evaluation of the methods chosen to develop the model was evaluated to interpret the knowledge patterns properly. The performance of each model developed in the study experiments are measured using accuracy, true

positive rate, false positive rate, precision, ROC area and 10 folds cross validation to evaluate the accuracy of the developed model.

3.3.6. Use the Discovered knowledge

The outcomes of the findings in this study would be reported, disseminated to and used by the concerned healthcare stakeholders and other interested practitioners. Therefore, interested domain experts and researchers can get access to the research results so as to support the decision-making process, or use it for advance research in the area or for any other applicable reasons.

CHAPTER FOUR

EXPERIMENTATION AND RESULT DISCUSSION

4.1. Overview

According to the methodology of this study after preparation of the data, the next task is the mining process. The main objective of this study is, to develop a model for predict the status of anemia among children aged 6 to 59 months using data mining techniques. Having this purpose in mind, the researcher has done model-building, which is carried out by using classification data mining approach.

This chapter also describes the results of the modeling and performance evaluation phases are presented in four sections. The first section presents the different experimentation using J48, Naïve Bayes and PART rule induction to build the models with all 20 attributes and selected 12 attributes as indicated the following sections. The second section presents the performance comparisons of the models. The comparison and evaluation were executed based on the results of the experiment performed using standard performance evaluation metrics. Here, the analysis and discussion of the output in the context of the problem were also presented. In the third part, we introduce rules extracted from the best model. Lastly, we present the potential set of attributes which can reasonably predict child with anemia instances.

In this study, during all experiments in order to validate and compare the classification performance of the techniques the 10-fold cross validation with default value was used. For all experiment, all the default parameter settings already offered by the WEKA tool were used, except for the parameter unpruned which had a default value False this means pruned and was changed to True for J48 and PART Rule induction method in order to examine the performance of the models.

Table4.1: Experiments and scenarios of the study

Algorisms	Parameters	Experiments	Scenario /models	Experments number
J48/ Decision tree	Pruned	10-fold cross validation with all attributes	1	1
		10-fold cross validation with selected attributes	2	
	Un pruned	10-fold cross validation with all attributes	1	2
		10-fold cross validation with selected attributes	2	
Naïve Bayes	Default	10-fold cross validation with all attributes	1	3
		10-fold cross validation with selected attributes	2	
PART/ Rule induction	Pruned	10-fold cross validation with all attributes	1	4
		10-fold cross validation with selected attributes	2	
	Un pruned	10-fold cross validation with all attributes	1	5
		10-fold cross validation with selected attributes	2	

4.2. Model Building

In this section, the results of the models built using the algorithms used in this study are presented. To build the predictive model, J48, naive Bayes and PART algorithms are

trained and evaluated. For training and testing the classification model the researcher used 10-fold cross validation methods the data was divided into 10 folds. Basically, five experiments were conducted and as a result a total of 10 models were developed. The experiments were conducted using all 20 attributes, and selected 12 attributes. To select such 12 attributes, the researcher used the gain ratio measure which was implemented by WEKA attribute ranking filter. Ranking the attributes basically indicates the relative importance of each input attribute in making a prediction (Witten and Frank, 2005). In this regard, the 12 selected attributes using the WEKA ranker filter were shown in Table 4.2.

Table4. 2: The Selected 12 attributes using WEKA ranker filter

No	Attribute name	Description
1	HHNUM	Number of household members
2	CHILD AGE	Current age of child
3	BIRTH INTERVAL	birth interval(months) of the child
4	WEALTH	Family wealth index
5	MEDUCLEV	Maternal education level
6	MOCCUP	Maternal occupation
7	MANEM	Maternal anemia level
8	NUMANV	Number of antenatal visit(during pregnancy)
9	RESIDENCE	Type of place of residence
10	WHZ	Child nutritional status
11	REGION	Region where the survey conducted
12	HAD DIARRHEA RECENTLY	Recent episodes of diarrhea

4.2.1. Model building using J48 decision tree

In this experiment, the predictive ability of J48 classifier is investigated. To build the models, two experiments considering four scenarios were conducted. The two scenarios

contained all 20 input attributes and the other four contained the selected 12 attributes. The results obtained from J48 classifier with all and selected attributes were summarized in Table 4.3 with their respective performance measures.

Table4. 3: J48 decision tree experiment results by applying all and best selected attributes

Experiment	Model	Detailed performance measures							
		Speed (sec)	TP Rate	FP Rate	Precision	Recall	F- Measures	ROC Area	Accuracy
I	J48 Pruned with all(20)attributes	0.11	0.930	0.095	0.917	0.930	0.923	0.942	91.805%
	J48 Pruned with selected(12)attributes	0.07	0.900	0.130	0.887	0.900	0.893	0.928	88.585%
II	J48 Unpruned with all(20)attributes	0.07	0.912	0.113	0.902	0.912	0.893	0.917	90.05%
	J48 Unpruned with Selected(12)attributes	0.05	0.892	0.130	0.886	0.892	0.889	0.914	88.190%

As we have seen from table 4.3 two experiments were executed. The first experiment was designed to evaluate the performance of J48 pruned model with all and selected attributes in predicting anemic children aged 6 to 59 months, whereas the second experiment was designed to evaluate the performance of J48 unpruned model with all and selected attributes in predicting anemic children aged 6 to 59 months. In each experiment, two scenarios/models were considered and the first scenario of the both experiments contained all attributes and the second scenario contained selected attributes.

In the first scenario of each experiment and second scenario of each experiment, the model was run on a full training set containing 8603 instances with all 20 attributes and

8603 instances and 12 selected attributes respectively. Using J48 decision tree classifier, various results were acquired during the above experiments. The first experiment/J48 pruned with all and selected attribute/ displays that the model built using J48 pruned by applying all (20) attributes 7898 (91.805%) instances are correctly predicted from 8603 instances, while the remaining 705 (8.194%) instances are incorrectly or falsely predicted and the result of J48 pruned with selected attributes shows that the model correctly classified 7621 (88.585%) instances, while 982 (11.414%) of the instances were classified incorrectly. This shows additional 277 instances are incorrectly predicted by J48 classifier when applying best selected attributes of the dataset in the classifier.

The second experiment shows that the model built using J48 unprund with all (20) attributes 7747 (90.05%) instances are correctly predicted from 8603 instances, while the remaining 856 (9.95%) instances are incorrectly or falsely predicted and the result of J48 unpruned with selected attributes shows that the model correctly classified 7587 (88.190%) instances, while 1016(11.809%) of the instances were classified incorrectly. From the above experiment, we have shown that considering the models built in each experiment, model‘s performance in both the pruned and unpruned scenarios with all attributes shown a better performance than when using selected attributes with their particular detailed performance measures except the speed. However, the pruned models performed better accuracy than the unpruned. The experimental outputs of the J48 pruned model with all attributes is presented in Appendix 2.

The base for calculating correctly classified instances and incorrectly classified instances is the confusion matrix. The confusion matrix of the class which is a base for calculating accuracy measures and performance is presented below. The output summary of J48 decision tree is shown in Tables 4.4 and 4.5

Table4.4: Confusion matrix output of experiment 1

		Actual class	Predicted class	
			Anemic	Not Anemic
Model 1	J48 pruned with all attributes	Anemic	4245(92.97%)	321(7.03%)
		Not anemic	384(9.51%)	3653((90.49%)
Model 2	J48 pruned with selected attributes	Anemic	4109(89.99%0	457(10.00)
		Not Anemic	525(13%)	3512(87%)

As we have seen in table 4.4, of the first model the number of True-positives correctly classified instances are 4245(92.97%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 321(7.03%) child were incorrectly classified as not anemic while in fact they belong to anemic. This model also correctly classified/True-negative/ 3653 (90.49%) child not anemic and the remaining 384(9.51%) child were incorrectly classified as anemic which should be categorized as not anemic. The second model the number of True-positives correctly classified 4109(89.99%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 457(10%) child were incorrectly classified as not anemic while in fact belong to anemic. This model also correctly classified/True-negative/ 3512(87%) child not anemic and the remaining 525(13%) child were incorrectly classified as anemic which should be categorized as not anemic. Generally, the sum of the True-positive and True-negative give us correctly classified instances. For example, J48 pruned with all attributes the total number of records which were correctly classified to “anemic and not anemic” classes were 7898 (91.805%) while 705(8.19%) records are incorrectly classified.

Table4.5: Confusion matrix of J48 decision tree with unpruned for experiment 2

		Actual class	Predicted class	
			Anemic	Not Anemic
Model 3	J48 Unpruned with all attributes	Anemic	4165(91.21%)	401(8.78%)
		Not anemic	455(11.27%)	3582((88.73%)
Model 4	J48 Unpruned with selected attributes	Anemic	4075(89.24%0	491(10.75)
		Not Anemic	525(13%)	3512(87%)

From the result in the above experiment model three correctly classified instances are 4165(91.21%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 401(8.78%) child were incorrectly classified as not anemic but in fact they belong to anemic. This model also correctly classified/True-negative/ 3582(88.73%) child not anemic and the remaining 455(11.27%) child were incorrectly classified as anemic which should be categorized as not anemic. The fourth model the number of True-positives correctly classified 4075(89.24%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 491(10.75%) child were incorrectly classified as not anemic but in fact belong to anemic. This model also correctly classified/True-negative/ 3512(87%) child not anemic and the remaining 525(13%) child were incorrectly classified as anemic which should be categorized as not anemic.

As we have seen in experiment 1 and 2, it is also possible to say that the J48 pruned model with all attributes is considered as the best model in terms of minimized incorrectly classified instances.

4.2.2. **Model Building Using Naïve Bayes Algorithm**

The second data mining technique used in this study was Naïve Bayes Algorithm with default parameter. To build the models, one experiment considering two scenarios was conducted. The first scenarios contained all 20 input attributes and the second one contained the selected 12 attributes. The results obtained from Naïve Bayes classifier with all and selected attributes were summarized in Table 4.6 with their respective performance measures.

Table4. 6: Naïve Bayes experiment results by applying all and best selected attributes

Experiment	Model	Detailed performance measures							
		Speed (sec)	TP Rate	FP Rate	Precision	Recall	F- Measures	ROC Area	Accuracy
I	Naïve Bayes with all(20)attributes	0.03	0.854	0.141	0.873	0.854	0.863	0.931	85.63
	Naïve Bayes with selected(12)attributes	0	0.856	0.147	0.868	0.856	0.862	0.931	85.47

As we have seen from table 4.6 one experiment were executed. This experiment was designed to evaluate the performance of Naïve Bayes model with all and selected attributes in predicting anemic child aged 6 to 59 months. This experiment, two scenarios/models were considered and the first scenario of the experiment contained all attributes and the second scenario contained selected attributes.

In the first scenario and second scenario of the experiment the model was run on a full training set containing 8603 instances with all 20 attributes and 8603 instances and 12 selected attributes respectively. In scenario one and two time taken to build model 0 second.

Using Naïve Bayes classifier, various results were acquired during the experiments. The first scenario with all attribute displays that the model built using Naïve Bayes classifier by applying all (20) attributes 7367 (85.633%) instances are correctly predicted from 8603 instances, while the remaining 1236 (14.367%) instances are incorrectly or falsely predicted and the second scenario with selected attributes shows that the model correctly classified 7353(85.47%) instances, while 1250(12.53%) of the instances were classified incorrectly. This shows additional 14 instances are incorrectly predicted by Naïve Bayes classifier when applying best selected attributes of the dataset in the classifier but the selected attributes of TP Rate, Recall and FP Rate performance measures are greater than all attributes whereas ROC Area of both attributes are equal.

However, the Naïve Bayes with all attributes models performed better accuracy compared to Naïve Bayes with selected attributes. The experimental results of the Naïve Bayes pruned model with all attributes is presented in APPENDEX 3.

Concerning the confusion matrix of the Naïve Bayes classifier technique, the output summaries are presented in Tables 4.7.

Table4. 7: Confusion Matrix Output of Naïve Bayes experiment 1

Model 1	Naïve Bayes with all attributes	Actual class	Predicted class	
			Anemic	Not Anemic
		Anemic	3898(85.37%)	668(14.63%)
		Not anemic	568(14.07%)	3469(85.93%)
Model 2	Naïve Bayes with selected attributes	Anemic	3910(85.63%)	656(14.37%)
		Not Anemic	594(14.71%)	3443(85.29%)

As we have seen in table 4.7 of the first model the number of True-positives correctly classified instances are 3898(85.37%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 321(14.63%) child were incorrectly classified as not anemic

while in fact they belong to anemic. This model also correctly classified/True-negative/ 3469 (85.93%) child not anemic and the remaining 568(14.07%) child were incorrectly classified as anemic which should be categorized as not anemic. The second model the number of True-positive correctly classified 3910(85.63%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 656(14.37%) child were incorrectly classified as not anemic while in fact belong to anemic. This model also correctly classified/True-negative/ 3443(85.29%) child not anemic and the remaining 594(14.71%) child were incorrectly classified as anemic which should be categorized as not anemic.

From these output, the Naïve Bayes pruned model with all attributes is the best model in terms of minimized incorrectly classified instances.

4.2.3. Model Building Using PART rule induction

The third data mining technique used in this study was PART rule induction classifier. To build this model, like J48 decision tree, two experiments were conducted and four scenarios were considered, two scenarios containing all 20 input attributes and the other two scenarios containing selected 12 attributes. The outputs obtained from PART rule induction classifier with all and selected attributes were summarized in Table 4.8 with their respective performance measures.

Table4. 8: PART rule induction experiment results by applying all and best selected

Experiment	Model	Detailed performance measures							
		Speed (sec)	TP Rate	FP Rate	Precision	Recall	F- Measures	ROC Area	Accuracy
I	PART Pruned with all(20)attributes	0.72	0.914	0.110	0.904	0.914	0.909	0.936	90.28%
	PART Pruned with	0.54	0.894	0.129	0.887	0.894	0.891	0.936	88.364%

	selected(12)attributes								
II	PART Unpruned with all(20)attributes	1.78	0.897	0.119	0.895	0.897	0.896	0.897	88.92%
	PART Unpruned with Selected(12)attributes	2.25	0.895	0.158	0.865	0.895	0.880	0.901	87%

As we have seen from table 4.8 two experiments were executed. The first experiment was designed to evaluate the performance PART pruned model with all and selected attributes in predicting anemic child aged 6 to 59 months, whereas the second experiment was designed to evaluate the performance of PART Unpruned model with all and selected attributes in predicting anemic child aged 6 to 59 months. In each experiment, two scenarios/models were considered and the first scenario of both experiments contained all attributes and the second scenario of each experiment contained selected attributes.

In the first scenario of each experiment and second scenario of each experiment the model was run on a full training set containing 8603 instances with all 20 attributes and 8603 instances and 12 selected attributes respectively. Using PART rule induction classifier, various results were acquired during the above experiments. PART pruned with all and selected attribute of the first experiment displays that the model built using PART pruned by applying all (20) attributes 7767(90.282%) instances are correctly predicted from 8603 instances, while the remaining 836(9.715%) instances are incorrectly or falsely predicted and the result of PART pruned with selected attributes shows that the model correctly classified 7602 (88.36%) instances, while 1001(11.64%) of the instances were classified incorrectly. This shows additional 165 instances are incorrectly predicted by PART classifier when applying best selected attributes of the dataset in the classifier.

The second experiment shows that the model built using PART unpruned with all (20) attributes 7650(88.922%) instances are correctly predicted from 8603 instances, while the remaining 953(11.078%) instances are incorrectly or falsely predicted and the result of

PART unpruned with selected attributes shows that the model correctly classified 7485 (87%) instances, while 1118(13%) of the instances were classified incorrectly.

From the above experiment, we have shown that considering the models built in each experiment, model‘s performance in both the pruned and unpruned scenarios with all attributes shown a better performance than when using selected attributes with their particular detailed performance measures except the speed. However, the PART pruned models with all attributes performed better accuracy compared to other models of PART rule induction. The experimental outputs of the PART pruned model with all attributes is presented in APPENDEX 4.

Concerning the confusion matrix of the PART rule induction classifier technique, the output summaries are presented in Tables 4.9 and 4.10

Table4. 9: Confusion Matrix of PART rule induction of Experiment 1

		Actual class	Predicted class	
			Anemic	Not Anemic
Model 1	PART pruned with all attributes	Anemic	4175(91.44%)	391(8.56%)
		Not anemic	445(11.023%)	3592(88.978%)
Model 2	PART pruned with selected attributes	Anemic	4084(89.44%)	482(10.56%)
		Not Anemic	519(12.86%)	3518(87.14%)

As we have seen in table 4.9, of the first model the number of True-positives correctly classified instances are 4175(91.44%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 391(8.56%) child were incorrectly classified as not anemic while in fact they belong to anemic. This model also correctly classified/True-negative/

3592 (88.978%) child not anemic and the remaining 445(11.023%) child were incorrectly classified as anemic which should be categorized as not anemic. The second model the number of True-positives correctly classified 4084(89.44%) child out of 4566 child aged 6 to 59 months as anemic and the remaining 482(10.56%) child were incorrectly classified as not anemic while in fact belong to anemic. This model also correctly classified/True-negative/ 3518 (87.14%) child not anemic and the remaining 519 (12.86%) child were incorrectly classified as anemic which should be categorized as not anemic.

Table4. 10: Confusion Matrix of PART rule induction of Experiment 2

		Actual class	Predicted class	
			Anemic	Not Anemic
Model 3	PART Unpruned with all attributes	Anemic	4094(86.66%)	472(10.34%)
		Not anemic	481(11.91%)	3556(88.09%)
Model 4	PART Unpruned with selected attributes	Anemic	4086(89.49%)	480(10.51%)
		Not Anemic	638(15.80%)	3399(84.2%)

As we have seen in table 4.10, of the first model the number of True-positives correctly classified instances are 4094(86.66%) child aged 6 to 59 months as anemic and the remaining 472(10.34%) child were incorrectly classified as not anemic while in fact they belong to anemic. This model also correctly classified/True-negative/ 3556(88.09%) child not anemic and the remaining 481(11.91%) child were incorrectly classified as anemic which should be categorized as not anemic. The second model the number of True-positives correctly classified 4086(89.49%) child aged 6 to 59 months as anemic and the remaining 480(10.51%) child were incorrectly classified as not anemic while in fact belong to anemic. This model also correctly classified/True-negative/ 3399 (84.2%)

child not anemic and the remaining 638 (15.8%) child were incorrectly classified as anemic which should be categorized as not anemic.

Finally from PART Rule confusion matrix, the PART pruned model with all attributes is the best model in terms of minimized incorrectly classified instances.

4.3.Empirical Analysis of Experiments

To develop the proposed predictive models, J48, Naïve Bayes and PART Rule induction algorithms are applied on the datasets. In the given datasets, a total of 10 models are developed by using all and best selected attributes. Gain ratio attributes evaluator with ranker search method is used to select the best and top ranked attributes for model development. In the data set five different experiments are conducted using the three classification algorithms, and tested using 10-fold cross validation evaluation method, which are chosen by the researcher for the conducted experimental research work. The experiments are considered to examine which algorithm is greatly accurate on developing predictive models. The study also checks the influences of attribute numbers on the performance of algorithms and assesses the effect of top ranked attribute selection on the overall performance of the algorithms. We showed that, on all experiments the prediction accuracy of all algorithms was different. In this investigation change on performances of algorithms are shown by applying attribute subset selection.

4.3.1. The Effect of Attribute Subset Selection summery

Attribute subset selection does not improve the performance of all models. Table 4.11 shows the accuracy variations of a J48, Naïve Bayes and PART Rule induction classifier before and after attribute subset selection is done using gain ratio attribute evaluator.

Table4. 11: J48**,** Naïve Bayes and PART Rule induction classifier accuracy variation on attribute subset selection

Model	Accuracy with all attributes	Accuracy with selected attributes	Changes
J48 pruned	91.805%	88.585%	-3.22
J48 unpruned	90.05%	88.19%	-186
PART pruned	90.28%	88.364%	-1.916
PART unpruned	88.92%	87%	-1.92
Naïve Bayes(default parameter)	88.63%	85.47%	-3.16

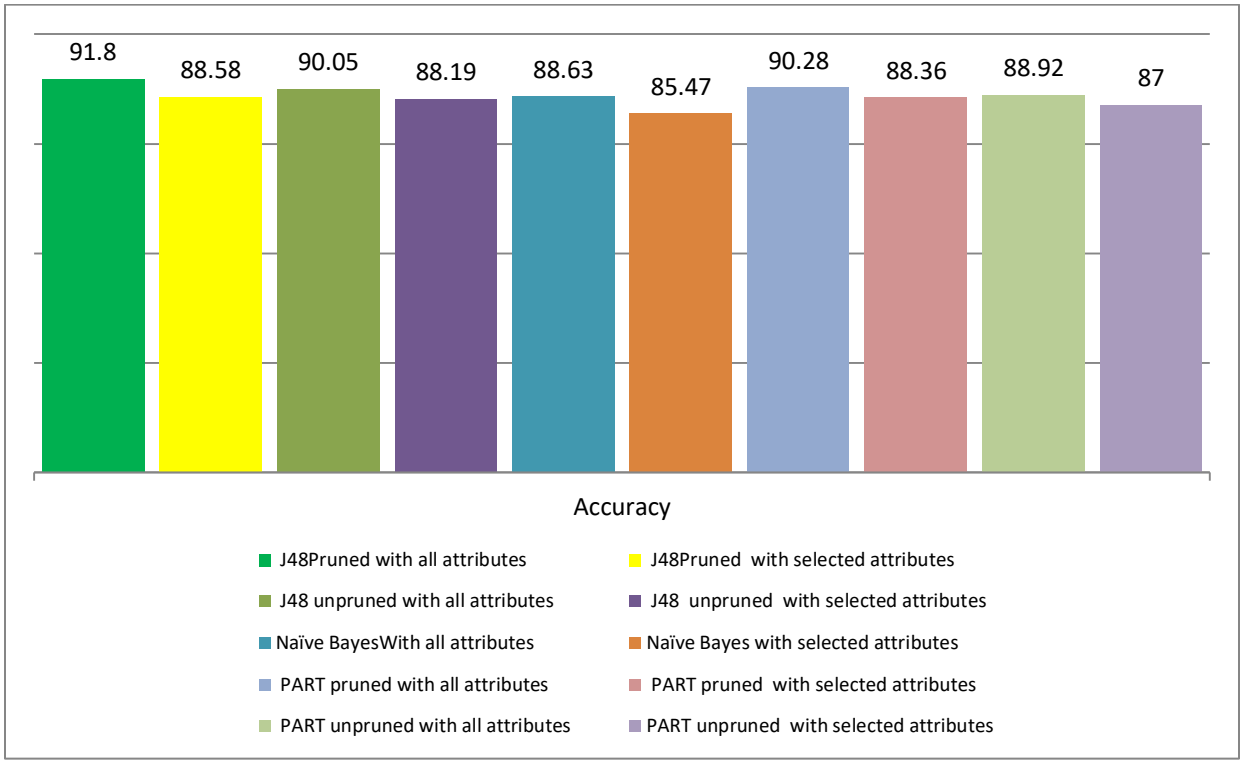

Figure4.1: Accuracy of classifiers

The performance of J48 and PART classifier with pruned and unpruned parameter has not shown any improvement, whether attributes subset selection is done on the dataset. Rather, its accuracy, TP rate, Precision, recall, F-measure, and ROC area (in J48) are decreased. But its FP rate with pruned and unpruned parameter, ROC with unpruned parameter is increased whereas ROC with pruned no changes. However, the performance

of Naïve Bayes classifier has shown some improvement, on TP, FP and recall. But equal in ROC and less in accuracy and F- measure whether attributes subset selection is done.

4.4. **Performance Evaluation**

This stage was comparing and evaluating the performance of the models in order to select the best model which is able to predicting status of anemia in child 6 to 59 months.

To compare and evaluate the performance of the models, the standard performance measures such as Accuracy, TP Rate, TN Rate, Precision, Recall, and F-Measure were used. ROC area and speed were also other parameters used to compare model‘s performance. In general, the comparison was performed based on the results of the standard performance evaluation measures.

From table 4.3 up to 4.10 the accuracy of all models in the experiments are, varies between 85.47- 91.805%, this describes lower scores for Naïve Bayes with selected attributes and higher scores for J48 pruned model with all attributes. The precision values, the highest scores of 0.917 were recorded by J48 pruned with all attributes followed by PART pruned model with all attributes that scored 0.914. The least scores were scored by the Naïve Bayes model with selected attributes 0.868. The Recall and F-Measure results, Naïve Bayes model with selected attributes scored the least scores of 0.856 and 0.862 respectively, whereas the highest score scored by J48 pruned with all attributes of 0.93(Recall) and 0.923 F-Measure were presented. The ROC area values, the highest scores of 0.942 were recorded by J48 pruned with all attributes and the least scores were scored by the J48 unpruned model with selected attributes 0.897.

The next performance measures to compare the model presented on the above confusion matrix table TP Rate and TN Rate. From the output of confusion matrix the highest and the least score TP Rate were registered by J48 pruned with all attributes of 92.97% and Naïve Bayes with all attributes of 85.37% respectively. The highest and the lowest TN

Rate were registered by J48 pruned with all attributes of 90.49% and PART unpruned with all attributes of 84.2% respectively. As we have seen in the confusion matrix the TP Rate (Sensitivity) scores of the models are somewhat higher than TN Rate (Specificity) scores of the models except Naïve Bayes model with all attributes. This implies, all models in the experiment were better in predicting positive cases as compared to negative cases. Based on the time taken to build a model, the Naïve Bayes model with selected attributes and all input attributes scored the fastest execution time than the other models of the experiment; this implies the number of attributes affect the execution time. From the performance results of each model it is observed that the models built using all attributes scored the best performance as compared to models built using selected attributes. Attribute selection does not improve the performance of the models. This result was supported by Molalign Desalew, (2018) compared the effect of selected attributes on the model performance. The model built with all input attributes perform better that of the model with selected attributes for most of the standard performance measures.

In general, the experimental result confirms that, the J48 pruned model with all attributes and PART pruned model with all attributes better than the other models by scoring the highest Accuracy, TP Rate, TN Rate, Precision, Recall, F-Measure and ROC area. But the Naïve Bayes with selected attributes scored the quickest run time.

As a result, based on these performance results, the J48 pruned model with all attributes was selected as the best model for this study. The predictive performance accuracy of the selected model was 91.805%, the model incorrectly classified 8.195% of the instances to some other classes. The potential reason may be due to the nature of the dataset used in this study.

4.5. Generated rules from the experiment

After successive experiments were build and the best decision tree model was selected. The next steps were to generate rule. Rules are generated by using J48 decision tree classifiers based on their prediction accuracy. Therefore, the model developed using J48 pruned with all attributes are the best model for predicting the status of anemia child aged 6 to 59 months.

The following are some of the rules extracted from the J48 pruned with all attributes. Therefore those which cover more cases and have better accuracy are chosen. The following rules indicate the possible conditions in which a child could be classified in anemic. The generated rules were evaluated by domain experts. The rules are listed and discussed below.

Rule 1: IF household number = 6 or more AND Child Age= <=2 years AND Maternal Occupation = Not Working AND Birth Interval of the child =<=36 months THEN the class Anemic (1557.0/29.0).
This rule is interpreted as: household number >=6 and maternal occupation status is not working and child age <=2 years and Birth interval of the child <36 months, there is high probability of anemic disease for the child.

Rule 2: IF child age <=2years AND maternal occupation = working AND maternal education level= No education AND Birth Interval of the child= <=36 months THEN Anemic (435.0/21.0).
The rule stated that: child age less than 2 years, and maternal occupation status working, and maternal education level has no education, and Birth Interval of the child has less than 36 months, then anemic child will be expected.

Rule 3: IF Residence= Rural AND wealth= Poor AND Maternal education level= No education AND Household number= 6 or more AND WHZ= Wasted THEN the class Anemic (150.0/6.0)

The rule is interpreted as: residence is rural, wealth status poor, maternal education no education, household number 6 or more and nutritional status waster then anemic child will be expected.

Rule 4: IF Household number= 6 or more AND child age <=2 years AND postnatal care of the child = <3 AND had diarrhea recently= Yes AND WHZ= Wasted THEN the class Anemic (54.0/5.0).

The rule is interpreted as: house hold number 6 or more, child age less than or equal to 2 years, PNC less than 3, had recently diarrhea recently yes and nutritional status wasted then anemic child will be expected

Rule 5: IF Child age = <=2 years AND Birth interval= <36 months AND household number= between 2 and 5 AND Maternal education level= No education THEN the class Anemic (402.0/31.0).

The fifth rule stated that child age less than or equal to 2 years, birth interval less than or equal to 36 months, household number between 2 and 5 and maternal education level no education is most likely child will be anemic.

Rule 6: IF Household number= 6 or more AND child age <=2years AND maternal occupation = working AND maternal education level= primary AND Breast feed status= No THEN the class Anemic (60.0/2.0)

The sixth rule stated that a child age less than or equal to 2 years, household number 6 or more, maternal occupation working, maternal education level primary, breast feeding status no then the child will be anemic.

Rule 7: IF household number= between 2 and 5 AND Child age = <2 years AND Birth Interval of the child <36 months AND Maternal education level= primary AND postnatal care of the child = >=3 AND residence= rural THEN the class Anemic (26.0/1.0).

The seventh rule stated that household number between 2 and five a child age less than or equal to 2 years, birth interval of the child less than 36 months, maternal education level primary, postnatal of the child greater than or equal to 3 and residence is equal to rural then the child will be anemic.

Rule 8: IF Child age = >=2 years AND Birth Interval of the child >=36 months AND PNC >=3 AND household number= between 2 and 5 THEN the class Not anemic (378.0/2.0).

The *eighth* rule stated that child age greater than equal to 2 years, birth interval of the child greater than or equal to 36 months, PNC greater than or equal 3 and household number between 2 and five a child age then the child will be not anemic.

Rule 9: IF household number= between 2 and 5 AND Child age >=2 years AND Birth Interval of the child >36 months AND vitamin A1 = no THEN the class Not anemic (926.0/48.0).

The ninth rule stated that household number between 2 and five a child age greater than or equal to 2 years, birth interval of the child greater than 36 months, vitamin A1 most recently received no then the child will be not anemic.

Rule 10: IF Household number= 6 or more AND child age <=2years AND maternal education level= no education AND Vitamin A1 received most recent = No AND birth interval of the child= <36 months THEN the class Anemic (673.0/63.0)

The tenth rule stated that household number six or more a child age less than or equal to 2 years, maternal education level no education, Vitamin A1 received most recent No AND birth interval of the childless than or equal to 36 months THEN the class Anemic

Rule 11: IF child age <=2years AND maternal education level= no education AND Vitamin A1 received most recent = No AND birth interval <36 months AND postnatal care of the child= <3 AND wealth = poor THEN the class Anemic (106.0/10.0)

The eleventh rule stated that child age less than or equal to 2 years, maternal education level no education, vitamin A1 received most recently no birth interval of the child less

than 36 months, postnatal care of the child less than three and wealth status of the family poor then the child will be anemic.

Rule 12: IF Household number= 6 or more AND child age <=2 years AND maternal occupation= Not working AND Postnatal care of the child = <3 THEN the class Anemic (422.0./29.0)

This rule states that household number six or more, child age less than or equal to two years, maternal occupation is not working, postnatal care of the child less than or three then the child will be anemic.

Rule 13: IF residence = rural AND wealth = poor AND child age <=2years AND maternal education level= no education AND household number= 6 or more AND number of antenatal visits = <4 AND WHZ = wasted THEN the class Anemic (160.0/2.0).

The last rule states that residence of the household is rural, wealth stats of the household is poor, child age less than or equals to 2 years, maternal education level no education, household number six or more, number of antenatal visit less than four and nutritional status of the child is wasted then the child will be anemic.

.

4.6. The potential set of attributes

One of the objectives of this study was to determine the potential set of attributes which can reasonably predict anemic children aged 6 to 59 months. In this research, 20 attributes were used for building a model. To discover the potential set of attributes from 20 attributes, the researcher used the results of the rule extracted from the J48 Pruned model with all attribute. The potential set of attributes are residence, wealth, WHZ(nutrition status), maternal education level, maternal occupation, Had recently diarrhea, household number, birth interval of the child, child age, postnatal care, received vitamin A1 most recent and current breast feed status, were consider as potentially the most determinant factors to predict any anemic children aged 6 to 59 months.

To verify these attributes as potentially determinant factors, the researcher performed experiments with the discovered 12 potential set of attributes. The outputs gained from 12 potential set of experiments were presented with their performance measures in Table 4.12.

Table4.12: Experiments with 12 potential set of attributes

Experiment	Model	Performance measures
		Accuracy
1	J48 pruned with potential set of attributes	90.224%
2	J48 unpruned with potential set of attributes	89.457%
3	Naïve Bayes with potential set of attributes	86.318%
4	PART pruned with potential set of attributes	90.084
5	PART unpruned with potential set of attributes	88.097

As it is shown in table 4.12 the output of the experiment proves that the experiment with 12 potential set of attributes better accuracy than all models with selected attributes. This makes sure that the identified attributes are potentially the determinant risk factors for anemia child aged 6 to 59 months.

Having those results with potential set of attributes then, the researcher consulted domain experts particularly nutritionist, and some previous published further works for other evidence. Whereas the field experts understood that the left behind attributes had also a role in prediction, domain experts agreed on such attributes which were identified as potential factors of anemia child aged 6 to 59 months. Furthermore, some previous studies, although reported on different patterns, also make sure these attributes as determinant risk factors for anemia in child aged 6 to 59 months. In this regard, the EDHS, 2016 report, rural residence, and mothers have no education, household wealth status and children's age as determinant factors for the prevalence of anemia in child aged 6 to 59 months. Anemia is more common in children from the poorest household,

those whose mothers have no education, children in rural areas and children whose age is less than 24 months.

A study by Rosa M.L. *et al* (2014), Child's age (less than 24 months), recent episodes of diarrhea, and those worse household conditions were the variables that most contributed to the prevalence of anemia. Other studies conducted by Dereje Habte, et al (2013) on Maternal Risk Factors for Childhood Anemia conclude mother's low education background and poor household economic status where found to be associated with low hemoglobin level in children. Studies conducted by simbauranga et al. (2015), on Prevalence and factors associated with severe anemia amongst under-five children", hospitalized at Bugando Medical Centre conclude children whose parents or caretakers working status are not working have a high risk of having anaemia. The experimental output of the J48 pruned with potential set of is presented in APPENDEX 5.

4.7. Error rate of the selected model

In classification or prediction techniques, the accuracy of the resulting model is measured either in terms of the percentage of instances correctly classified or in terms of "error rate" i.e. the percentage of records incorrectly classified.

The classification error rate for the selected model is 8.195%, which means the model has incorrectly classified about 8.195% instances out of their actual classes each time when the model is tested on the test set. The percentage of incorrectly classified instances indicates the chance with which the developed model misclassifies a new victim out of the actual class. Several reasons may be predictable for increased error rate/ incorrectly classified instances/ from the models. First, attributes may not be included in the collection and study might have influenced it. Second, algorithms differ in their capability as observed from comparisons of performance measures. The other reason for misclassification is due to the fact that children anemia status (anemic or not anemic) is

based on the values of other attributes i.e. taking the similarity of the other attributes as predominant predictive values. All the models of the predictive performance in identifying True Positive cases of model are higher than identifying True Negative cases. This is because there is imbalanced between the two classes in the dataset therefore; the model tends to misclassify instances to some other classes.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as healthcare sector, has been applied for patient survival analysis, prediction of diagnosis, for outcomes measurement, to improve patient care and decision-making, etc. However, the potentials of data mining have not yet been used for predicting the status of anemia children aged 6 to 59 months in Ethiopia. The purpose of this study was to build a predictive model for the status of anemia among children aged 6 to 59 months using data mining techniques. The researcher used EDHS 2016 data collected by the Central Statistics Agency (CSA) as a source for this study.

The study implemented the hybrid, iterative methodology, was used, which consists of six basic steps such as understanding the problem domain, data understanding, data preparation, data mining, and evaluation of the discovered knowledge and use of the discovered knowledge. In order to generate interesting rule from the huge data collected in the 2016 EDHS data set, a total of 8,603 instances out of 41,392 instances and 20 independent attributes out of 1245 variables were applied. The independent variables/attributes were: maternal's age, region, residence, maternal's occupation, maternal's education level, maternal's BMI, wealth index, number of household, age of child, sex of child, size of child at birth, birth interval, maternal anemia level, WHZ, father's education level, Antenatal care during pregnancy, postnatal care of a child, Received Vitamin A1 (most recent), Currently breastfeeding status, had recently diarrhea and dependent attribute was children anemic level.

The experiment is done using three popular data mining, classification algorithms such J48 decision tree, Naïve Bayes and PART rule induction. 10 - Fold Cross Validation is adopted as a test option for random sampling of the training and test data samples.

Several models were built during experimentation that could predict the status of anemia. Among these models, J48 pruned model with all attributes outperformed the other models by achieving the highest Accuracy, TP Rate, TN Rate, Precision, Recall, FMeasure, and ROC area. The highest predictive accuracy was scored by J48 pruned model with all attributes 91.805% while Naïve Bayes with selected attributes scored the least accuracy of 85.47%. The outcomes clearly suggested that most of the independent attributes had strong relation with anemia status of children aged 6 to 59 months in the demographic and health survey data.

In summary residence, wealth, WHZ(nutrition status), maternal education level, maternal occupation, Had recently diarrhea, household number, birth interval of the child, child age, postnatal care, received vitamin A1 most recent and current breast feed status, were consider as the potential set of attributes to predict the status of any anemic children aged 6 to 59 months in Ethiopia.

5.2. Recommendation

In this research work, efforts have been made to apply data mining technology in predicting anemia status of children aged 6 to 59 months based on demographic, health and socio-economic characteristics.

Thus, based on the result of the research, the researcher would like to make the following recommendations. Reducing anemia diseases and improving child health status, through appropriate involvements, requires better understanding of the main demographic, health

and socioeconomic determinants. Thus, based on the result of the research learned by J48 with pruned algorithm; the following recommendations were made by the researchers.

- It has been observed that developing many other classifiers for prediction with the short period of time given to this research was unlikely. Therefore, to enhance the performance of the present model further research should be conducted on anemia status of children aged 6 to 59 months incrementally using many more mining techniques to improve the predictive model accuracy.

- The present study has considered demographic and health survey dataset to predict anemia status of children aged 6 to 59 months. Clinical data that have been gathered from different health care institutions should pay attention in anemia diseases. So that future study needs to discover knowledge and patters in clinical datasets and compare it with the result obtained using demographic and health survey dataset datasets.

- All the predictors managed through using different strategy such as giving health education on anemia, childcare at community based through rural health extension workers at rural area will minimize the anemia problem of children aged 6 to 59 months. Therefore, health professionals who are working on anemia should strengthen appropriate health involvement.

RFERENCES

A,Olani. (2008). "Predicting First year University Students' Academic Success," Institute for Educational Research.

Abraham T. (2005). Application of Data Mining technology to identify determinant risk

factors of HIV infection and to find their association rules: the case of Center for Disease Control and Prevention (CDC). Unpublished Master's Thesis Addis Ababa University, Addis Ababa.

Behailu Gebre Mariam and Tesfahun Hailemariam. (2012). Application of data mining for predicting adult mortality. [M.sc. thesis]. Addis Ababa University, Ethiopia.

Bentley, M and Griffiths, P. (2003). The burden of anemia among women in India. European Journal of Clinical Nutrition. [Cited Janwary 10, 2019].

Bereket Geze, Melese Sinaga and Tefera Belachew. (2018). Anemia and associated factors among children aged 6–23 months in Damot Sore District, Wolaita Zone, South Ethiopia.

Biset D. (2011). Predicting low birth weight using data mining techniques on Ethiopian Demographic and Health Survey data sets. [M.sc. thesis]. Addis Ababa University, Ethiopia.

Bresnahan J. (2000). Data Mining in the Health care: A Delicate operation. Available URL : <http://www.Cio.com/archive/061597-mining-content.html>. [Cited February 08, 2019].

Cao. L. (2007). *Fraud detection using Data mining*. Wiley IEEE Press, London.

Central Statistical Agency. (2017). Ethiopia and ICF International. Ethiopia Demographic and Health Survey. (2016). Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International.

Charles Patrick.Hemoglobin. (1996). (Low and High Range Causes). Medical Editor: URL: https://www.medicinenet.com/hemoglobin/article.html. [Cited August 10, 2018].

Cios Krzysztof J., Pedrycz Wiltod., Swiniarski Roman W. Kurgan Lukasz A. (2007). Data Mining: A knowledge Discovery approach. New York: Springer-Verlag Science Business Media LLC;

Daniel T. L. (2005). Discovering knowledge in data: An introduction to data mining. Canada: John Wiley & sons.

Deogan. (2011). Data Mining: research Trends, Challenges, and Applications [database

ontheInternet].http://citeseer.nj.nec.com/deogun97data.html [Accessed on february,21,2019].

Dereje Habte, Kalid Asrat, Mgaywa GMD Magafu, Ibrahim M. Ali, Tadele Benti, Wubeshet Abtew Girma Tegegne, Dereje Abera, Solomon Shiferaw. (2013). "*Maternal Risk Factors for Childhood Anemia in Ethiopia*", African Journal of Reproductive Health.

Ethiopia Central Statistical Agency. (2011). Ethiopia Demographic and Health Survey 2011. Calverton, MD: Addis Ababa Ethiopia: Central Statistics Agency. (Preliminary Report).

Ethiopia Central Statistical Agency. (2016). Ethiopia Demographic and Health Survey 2016. Addis Ababa Ethiopia: Central Statistics Agency;. (Preliminary Report).

Fayyad U, Piatetsky-shapiro, G. and Smyth, Padharic. From Data Mining to Knowledge Discovery in Databases. (1996). database on the Internet .[Cited April 15,2019]

F.T. (2006). "Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining," International Conference on Information Acquisition, vol. 1, no. 1.

Getachew Setotaw, june. (2013). predicting the status of anemia in women aged 15-49 by applying data mining techniques using the 2011 Ethiopia demographic and health survey (EDHS) datasets.

Getachew Mullu Kassa, Abadi Kidanemariam Berhe, Gedefaw Abeje Fekadu and Achenef Asmamaw Muche. (2017). Prevalence and determinants of anemia among pregnant women in Ethiopia. a systematic review and meta-analysis.

Guo Y, Grossman R. (1999). High performance data mining Scaling Algorithms, Applications and Systems.2002. USA: Kulwer Academic publisher.

Han J, Kamber, M. (2006). "Data Mining: Concepts and Techniques," Second Edition, Morgan Kauffman Publishers, San Francisco.

Han J, Kamber, M. (2012). Data mining: concepts and techniques, 3rd ed. San Fransisco, CA: Morgan Kaufmann,

Hung Y, McCullagh P, Black N, Harper R. Feature Selection and Classification Model Construction on Type 2 Diabetic Patient's Data. <http://citeseer.nj.nec.com/deogun97data.html> [Accessed on february,21,2019].

Jiawei al et. (2012). Data Mining Concepts and Techniques,3rd Edition. Waltham, USA: Morgan Kaufmann Publisher.

Kantardzic M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms. New York: John Wiley & Sons.

Kim, Y. Street, and Menczer, F. (2000). Feature selection in Data Mining. [Cited March 22, 2019]. University of Iowa, USA.

Mary, K. & Obenshain, M. (2004). Application of Data Mining Techniques to Healthcare Data. Chicago Journals The Society for Healthcare Epidemiology of America , pp. 690-695

M. Bharati, (2013) "Data Mining Techniques and Applications," *Indian Journal of Computer Science and Engineering*, vol. 1, no. 4, pp. 301-305.

Molalign desalew, (2018). *data mining based critical analysis and classification of infectious disease for effective diagnosis and treatment*. [Msc thesis]. Bahir Dar, Ethiopia. 2018

Muluneh Endalew. (2011). Exploring the prevalence of diarrheal disease using data mining technology: a case of Tikur Anbessa Hospital. [M.sc. thesis]. Addis Ababa University, Ethiopia..

Ng AYJ, M. I. (2002). On Discriminative vs. Generative Classifiers. A comparison of Logistic Regression and Naive Bayes, Neural Information Processing Systems.

O. Marban, G. Mariscal, and J. Segovia. (2009). "A Data Mining and Knowledge Discovery Process Model," I-Tech Education and Publishing.

P. Neelamadhab, M. Pragnyaban and P. Rasmita. (2012) "The Survey of Data Mining Applications and Feature Scope," *International Journal of Computer Science Engineering and Information Technology,* vol. 2, no. 3.

R. Agrawal. (2015)."Mining Association Rules between Sets of Items in Large Databases.," in In Proceedings of SIGMOD, 20716, 1993.According to (Sutch, T. (2015). Using association rules to understand subject choice at AS/A level., 2015.

R. Pressman. (2005)."Software Engineering: A Practitioner's Approach," vol. 1, no. 1.

Rodak T, Bernadette F. (2007). Hematology: clinical principles and applications (3rd Ed.). Philadelphia: Saunders; 220

Rosa M.L. Semedo1, Marta M.A.S. Santos1, Mirian R. Baião1, Ronir R. Luiz2, Gloria V. da Veiga. (2014). *Prevalence of Anaemia and Associated Factors among Children below Five Years of Age* in Cape Verde, West Africa

S. Fadzilah and A. Mansour. (2011). "Knowledge-Oriented Applications in Data Mining," InTech.

Senthilkumar D and Paulraj. (2015). "Prediction of Low Birth Weight Infants and Its Risk Factors Using Data Mining Techniques" International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates (UAE).

Sharman A. (2000). Anemia testing in population-based surveys: general information and guidelines for country monitors and program managers. Calverton, Maryland USA: ORC Macro.

Shegaw Anagaw. (2002). "Application of data mining technology to predict child mortality patterns",: the case of Butajira Rural Health Project (BRHP). [M.sc. thesis]. Addis Ababa University, Ethiopia.

Simbauranga et al. (2015). "Prevalence and factors associated with severe anaemia amongst under-five children", hospitalized at Bugando Medical Centre, Mwanza, Tanzania.

Stevens GA, Finucane MM, De- Regil LM, paciorek CJ, Flaxman SR, Branca F, et al. (2013). *Global, regional, and national trends in hemoglobin concentration and prevalence of total and severe anemia in children and preginant and non-preginant women* for 1995-2011: a systematic analysis of population representative data. Lancet glob health.

T. Sutch. (2015)."Using association rules to understand subject choice at AS/A level.," Cambridge,

Two Crows Corporation. (2005). "Introduction to Data Mining and Knowledge Discovery,".*3rd*. Potomac, MD 20854, USA.

Weiss, Sholom M. and Zhang, Tong. (2003).Performance analysis and evaluation. In:Ye Nong, editor. The Hand book of data mining. New Jeresy, USA: Lawerence Erlbaum Associates Inc.

WHO. (2008). *Worldwide prevalence of anemia 1993–2005*. WHO global database on anemia Geneva.

WHO. (2014). Global nutrition targets 2025: anemia policy brief (WHO/NMH/NHD/14.4). Geneva: World Health Organization.

WHO. (2015). *The global prevalence of anemia in 2011*. Geneva: World Health Organization.

Witten IHaF, Eibe. (2000). Practical Machine Learning Tools and Techniques with Java Implementations.USA: Academic Press.

Witten, I. and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann publishers.

Witten, I. H. and Frank, E. (2005). Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, CA, Morgan Kaufman.

World Health Organization. (1992). The Prevalence of anaemia in women: a tabulation of available information, 2nd ed.

APPENDEX 1**.** Name of Consulted Domain Experts

No	Name	Sex	Field	Institute
1	Habtamu G/mariam	m	MPH	Woldia health office
2	Migbar Woreta	m	Pharmacist	Woldia hospital
3	Tewodiros Molla	M	Nutritionists	Woldia

APPENDEX 2: J48 Pruned decision tree model with all attributes

=== Run information ===

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

Relation: Finaldataset

Instances: 8603

Attributes: 21

Region

Residence

Wealth

HHNUM

FEDUCLEV

MEDUCLEV

MAGE

MBMI

MOCCUP

MANEM

SEX

Child age

Birth interval

BREASTF

PNC

NUMANV

Child size

Had diarrhea recently

Vitamin A1

WHZ

CHILD ANEM

Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves: 183

Size of the tree : 283

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7898	91.8052 %
Incorrectly Classified Instances	705	8.1948 %
Total Number of Instances	8603	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.930	0.095	0.917	0.930	0.923	0.835	0.942	0.915	ANEMIC
0.905	0.070	0.919	0.905	0.912	0.835	0.942	0.935	NOT ANEMIC
0.918	0.083	0.918	0.918	0.918	0.835	0.942	0.924	Weighted Avg.

=== Confusion Matrix ===

a	b	<-- classified as
4245	321	a = ANEMIC
384	3653	b = NOT ANEMIC

APPENDEX 3: Naïve Bayes model with all attributes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: Finaldataset

Instances: 8603

Attributes: 21

Region

Residence

Wealth

HHNUM

FEDUCLEV

MEDUCLEV

MAGE

MBMI

MOCCUP

MANEM

SEX

Child age

Birth interval

BREASTF

PNC

NUMANV

Child size

Had diarrhea recently

Vitamin A1

WHZ

CHILD ANEM

Test mode: 10-fold cross-validation

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

Naïve Bayes Classifier

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 7367 85.6329%

Incorrectly Classified Instances 1236 14.3671%

Total Number of Instances 8603

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.854	0.141	0.873	0.854	0.863	0.712	0.931	0.940	ANEMIC
0.859	0.146	0.839	0.859	0.849	0.712	0.931	0.919	NOT ANEMIC
0.856	0.143	0.857	0.856	0.856	0.712	0.931	0.930	Weighted Avg.

=== Confusion Matrix ===

a	b	<-- classified as
3898	668	a = ANEMIC
568	3469	b = NOT ANEMIC

APPENDEX 4: PART pruned model with all attributes

=== Run information ===

Scheme: weka.classifiers.rules.PART –M 2 –C 0.25 –Q 1

Relation: Finaldataset

Instances: 8603

Attributes: 21

Region

Residence

Wealth

HHNUM

FEDUCLEV

MEDUCLEV

MAGE

MBMI

MOCCUP

MANEM

SEX

Child age

Birth interval

BREASTF

PNC

NUMANV

Child size

Had diarrhea recently

Vitamin A1

WHZ

CHILD ANEM

Test mode: 10-fold cross-validation

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

Naïve Bayes Classifier

Time taken to build model: 0.72 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7767	90.2825%
Incorrectly Classified Instances	836	9.7175%
Total Number of Instances	8603	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.914	0.110	0.904	0.914	0.909	0.805	0.936	0.908	ANEMIC
0.890	0.086	0.902	0.890	0.896	0.805	0.936	0.916	NOT ANEMIC
0.903	0.099	0.903	0.903	0.903	0.805	0.936	0.911	Weighted Avg.

=== Confusion Matrix ===

a	b	<-- classified as
4175	391	a = ANEMIC
445	3592	b = NOT ANEMIC

APPENDEX 5: J48 Pruned decision tree model with potential set of attributes

=== Run information ===

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

Relation: Finaldataset

Instances: 8603

Attributes: 13

Residence

Wealth

HHNUM

MEDUCLEV

MOCCUP

Child AGE

Birth interval

BREASTF

PNC

Had recently diarrhea

Vitamin A1

WHZ

CHILD ANEM

Test mode: 10-fold cross-validation

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

J48 decision list

Number of Leaves: 110

Size of the tree: 193

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7762	90.224%
Incorrectly Classified Instances	841	9.775%
Total Number of Instances	8603	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.908	0.105	0.908	0.9.8	0.908	0.804	0.946	0.932	ANEMIC
0.895	0.092	0.896	0.895	0.896	0.804	0.946	0.939	NOT ANEMIC
0.902	0.99	0.902	0.902	0.902	0.804	0.946	0.936	Weighted Avg

=== Confusion Matrix ===

a	b	<-- classified as
44147	419	a = ANEMIC
422	3615	b = NOT ANEMIC