ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

EXPLORING THE PREVALENCE OF DIARRHEAL DISEASE USING DATA MINING TECHNOLOGY

(A CASE OF TIKUR ANBESSA HOSPITAL)

By

Muluneh Endalew

June, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

EXPLORING THE PREVALENCE OF DIARRHEAL DISEASE USING DATA MINING TECHNOLOGY

(A CASE OF TIKUR ANBESSA HOSPITAL)

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Health Informatics

By

Muluneh Endalew

June,2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

EXPLORING THE PREVALENCE OF DIARRHEAL DISEASE USING DATA MINING TECHNOLOGY

(A CASE OF TIKUR ANBESSA HOSPITAL)

By

Muluneh Endalew

June, 2011

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor(s),	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner,	_____	

Declaration

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

_____

Date

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

_____

_____

Advisors

Acknowledgement

I wish to express my profound gratitude to my advisors Ato Ermias Abebe, and Dr. Assefa Seme for their constant support and constructive suggestions and overall guidance during my work. I am also grateful to Ato Benebere Mulugeta who is the chief of the center which I collected from the data for his professional inputs and support during my work

I am also grateful to the Staff of Pediatrics department who allowed access to this data to my work.

My special thanks are also due to all my friends for their all rounded contribution and support in the course of conducting this research work.

To all others who helped me in completing this study, I am equally grateful.

Table of Contents

List of Tables

List of Figures

Abbreviations

ANN=Artificial Neural Network

AWD=Acute Watery Diarrhea

CRISP-DM= CRoss-Industry Standard Process for Data Mining

DM=Data Mining

DT=Decision Tree

FMoH=Federal Ministry of Health

FN=False Negative

FP=False Positive

ICT=Information Communication Technology

KDD=Kowledge Discovery in Data bases

K-NN= K-nearest neighbor

NB=Naïve Bayes

ORS=Oral Rehydration Salt

RL=Ringer Lactate

SMOTE= Synthetic Minority Oversampling Technique

TN=True Negative

TP=True Positive

WHO=World Health Organization

ABSTRACT

The amount of health related data available to healthcare providing organizations for various diseases is being massive and ongoing to collect from time to time. As a result, huge amount of data is being stored in the health care organizations and facilities. Diarrheal disease is one of those which is being the causes of morbidity and mortality for many children especially under the age of five and from which large amount of data is being collected in both Rural and Urban health facilities of Ethiopia. This data represents a useful resource for making a wide variety of real-time decisions and determinations, from the quality of care delivered to trends in treatment modalities and staffing issues.

The problem is to be able to handle this huge amount of data and information in such a way that they can identify what is important and be able to extract it from the accumulated data. It is too complex and voluminous to be processed and analyzed by traditional methods. Now a days, data mining technology is being used as a tool that provides the techniques to transform these mounds of data into useful information which in turn enables to derive knowledge for decision making. A number of data mining techniques and tools are available to perform this task. The researcher considered selective techniques and tools which were used to explore the prevalence of diarrheal disease and develop classification and prediction models.

Thus, the purpose of this study is 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 center of African sub Region II in Tikur Anbessa Hospital. Patients' records with age of five years (60 months) and under are included in the study. Two machine learning algorithms from WEKA software such as J48 Decision Trees(DT) and Naïve Bayes(NB) classifiers are adopted to classify diarrheal disease records on the basis of the values of attributes 'Treatment' and 'Type of Diarrhea'. Initially, a total dataset of 5,572 records with 9 attributes were collected for the study. However, the size of class labels for the selected target classes was not balanced

and number of records were resample using ‘SMOTE (Synthetic Minority Oversampling TEchnique) from Weka preprocess package. After this process, the number of records used for model building was increased to 13,710 and 16, 460, for ‘Treatment’ and ‘Type of diarrhea’ target classes respectively. This was done in order to decrease biasness or preconception of classifiers in model building process.

Results of the experiments have shown that J48 DT classifier has better classification and accuracy performance as compared to NB classifier. Two consecutive models selected in evaluation performance of these classifiers depicted that J48 DT and NB classified ‘treatment modalities’ and ‘diarrheal types’ with the accuracy of 88.3%, 79.54%, 85.64% and 73.94% respectively. Overall, this study has proved that data mining techniques are valuable to support and scale up the efficacy of health care services provision process.

CHAPTER ONE

1. INTRODUCTION

1.1 Background

The outbreak and prevalence of various diseases in different times is a common phenomenon in the world specifically in developing countries. Among these, diarrhea is one of the severe diseases, especially for children under the age of five. It is a major cause of morbidity and mortality among children accounting for around three million deaths in developing countries per year. About 80% of these deaths occurred in the first two years of life. Across the globe there are an estimated 1.8 billion episodes of childhood diarrhea annually, mostly in developing countries (WHO, 1995).

Diarrheal disease in Ethiopia is also one of the major causes of morbidity and mortality for children under the age of five. It is estimated that 230,000 under five children die due to diarrhea and an average of five episodes of diarrhea occur per child per year (Damte et al., 2008). Poor personal and domestic hygiene, unsanitary excreta disposal, unsafe and inadequate water supply and malnutrition are the major risk factors among others (G.Mitikie, 2001). Thus, the number of people feeling sick and getting admitted to get treatment in health facilities related to this disease were increasing and it in turn raised the amount of data to be stored. As a result, Healthcare facilities have at their disposal vast amounts of data. As the data being stored becomes huge, analyzing, interpreting and making maximum use of it manually became difficult. This fact shows that we are getting more and more inundated by data/information and yet ravenous for knowledge.

In Ethiopia, the healthcare institutions on their daily activities were collecting and storing huge amount of data of which most of it was manual or paper based. But, the data in most

of these institutions have piled up on an unprecedented scale without any analysis. Among these, the diarrheal disease control and training center of African Sub Region II which is established by the Federal Ministry of Health (FMoH) and World Health Organization (WHO) twenty years ago in Tikur Anbessa hospital was storing data related to diarrheal disease. But, to the best knowledge of the researcher, the data stored in this center was not used by researchers specifically using recent ICT tools and techniques.

In the developed world, advances in research methods and new technologies made health care data gathering practices more electronic, dynamic and making it extremely rich for discovering knowledge. Besides, the inability of human beings to analyze the huge accumulated data and make use of them for decision-making has created a need for development of new technology and tools. (K.C, 2001). One of the technologies that yield exciting results for almost every organization that collects data on its customers, markets, products or processes is data mining. By discovering hidden patterns and relationships in the data, data mining enables users to extract greater value from their data than simple query and analysis approaches. Regardless of its relatively recent developed methodology and technology, it is being used by many organizations. Among these, in the health care sector, it is becoming increasingly popular, if not increasingly crucial (Birnbaum & Obenshain, 2004). This technology enables health care sectors to transform healthcare raw data into healthcare enterprise decisions. It also enables health care professionals and other concerned bodies to make clinical decisions and provide timely, cost effective and quality medical services (Koh & Tan, 2005).

Thus, the overall purpose of this study is 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 center of African sub Region II in Tikur Anbessa Hospital. The techniques used for this purpose were J48 DT (Decision Tree) and NB (Naïve Bayes) which are among the facilities comprised in Weka software.

1.2 Statement of the Problem

Human experience for the past years has shown that there is a great sorrow on the death of infants and children. The situation is worse in developing countries where among other things, infant and child mortality rates are high. This is mainly accompanied by factors such as the scarcity of services and facilities, poor sanitary conceptions, and improper utilization of the resources they have (Mirgissa & Fikau, 2000). Diarrheal disease is one of those which are accountable for such deaths. Worldwide, it is one of the major causes of morbidity and mortality for children in particular under the age of five. Most of the deaths are registered in the Africa region. Of the estimated total 10.6 million deaths among children younger than five years of age worldwide, 42 percent occur in African region (WHO, 2005).

Ethiopia is one of African countries which are highly affected by this disease. Various reports and studies have shown that diarrheal disease in Ethiopia is one of the major causes of mortality and morbidity next to lower respiratory tract infection specifically for children under the age of five (Damte et al.,2008). The proportion of mortality associated with diarrhea in Ethiopia is about 22.6% (A. Mekasha and A. Tesfahun, 2003). In this country, the prevalence of this disease is highly aggravated by poor sanitation, inadequate and unsafe water supply, poor personal and domestic hygiene and malnutrition. Ignorance, poor socio-economic status and low health service coverage are also other factors to the increased prevalence of diarrheal diseases (G.Mitikie, 2001). In Addis Ababa, the capital of the country and which has population of about 3.0 million people, diarrheal disease is one of the causes of morbidity and mortality. For example, in 2009,261 deaths were registered from June to August out of 901 cases (ERCS, 2009). The factors mentioned above are attributable for its outbreak and prevalence. As a result, the health care organization and facilities have been storing huge amount of clinical data whenever they deliver medical services related to this disease. The diarrheal disease control and training center of African sub Region II in Tikur Anbessa Hospital is one of

those health facilities that provide medical services and storing data on daily basis for diarrheal cases. Although large amount of patients' data related to the disease is being stored in different health care facilities including the aforementioned center, mostly, clinical decisions are still often made based on health care professionals' intuition and experience rather than on the knowledge rich data hidden in previously stored data. Digging of useful information and knowledge from the accumulated health care data for decision making is still not a common practice. But, successful decision systems enriched with analytical solutions are necessary for healthcare information systems. The delivery of health care is becoming data intensive, and there is a need to process and analyze this data in order to get insights about disease prevalence and the factors related to it.

A number of studies related to diarrhea disease were carried out in the country and the occurrence of diarrhea in children specifically under the age of five by different factors has been described. Various interventions were also pointed out in some of these studies. For example, a cross-sectional survey from Gondar, Northwest Ethiopia, indicated that the use of unprotected water sources was significantly associated with diarrheal morbidity. In this survey, a high prevalence of dysentery/bloody and persistent diarrhea were observed (G.Mitikie, 2001). Another cross-sectional Community Based Study from Urban South Western Ethiopia, Jimma, revealed that well source of water, lack of complete immunization, attack of measles and acute respiratory infections (ARI) were to be significantly associated with occurrence of diarrheal disease and persistent diarrhea was very high (A. Mekasha and A. Tesfahun, 2003). As an intervention, a study in Addis Ababa, Tikur Anbessa Hospital revealed that Zinc supplementation in the treatment of childhood diarrhea is well tolerated by patients (Damte et al., 2008). ORS as one of the major intervention mechanism for diarrhea was shown in another cross-sectional study, Mana District, Jimma Zone, South West Ethiopia (Mirgissa & Fikau, 2000).

Although various findings with all these studies mentioned above and other studies related to diarrhea have been disclosed, they were not beyond simple cross-sectional

survey of the disease in different locations and using simple statistical techniques to analyze the data rather than using advanced IT/ICT tools/ techniques to extract useful knowledge from the previous patients' stored data. Most practices in health care institutions that are not assisted by recent advanced technologies leads to unwanted errors, excessive medical costs and disastrous consequences which are therefore unacceptable and affects the quality of service provided to patients.

Therefore, the aim of this research is to investigate and demonstrate 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 center of African sub Region II in Tikur Anbessa Hospital.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of this research is 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 center of African sub Region II in Tikur Anbessa Hospital.

1.3.2 Specific Objectives

The specific objectives of the study include:

1. To assess the potential of data mining techniques in assisting health care related decisions;
2. To extract classification and prediction rules that can support to assess diarrheal disease prevalence ;
3. To Build and train models and test their performance;
4. To compare the classification and prediction performance of decision tree and Naïve Bayes data mining techniques for health care data;
5. To Report the result and forward recommendation.

To perform the above stated objectives, the following tasks were carried out by the researcher.

- Conducted a thorough review of literature on the data mining techniques and methods in general, and their application in the healthcare sector in particular;
- Identified data sources and collected required data from the specified center;
- prepared the data for analysis;
- Identified and selected appropriate data mining software and algorithms that were used in extracting meaningful patterns and relationships from the prepared data.

1.4 Scope and Limitations of the Study

The scope of this study was to investigate the potential applicability of classification and prediction techniques in exploring the prevalence of diarrheal disease. Though the tool selected for this study, which is Weka software, provides many classification and prediction techniques, this study was restricted to applying only J48 DT and NB techniques.

The scope of this research was also restricted to the data collected only from the diarrheal disease control and training center of African sub Region II in Tikur Anbessa hospital. The data which was stored from 1995-2000 E.C. was employed for the study. The age groups included in this study were those the age of five years and under (<=60 months).

The availability of accessible and usable data/databases for this research was one of the limitations that encountered the researcher to undertake this research. Because, the researcher was expected to enter the data from the paper based file in to the computer to make it appropriate for the study. Thus, the data conversion process, that is, data entry process to change the hard copy format of the data in to electronic format consumed much time. Hence, lack of sufficient time was another limitation of the study. Inadequacy of budget was also a limitation of the study.

1.5 Research Contribution

One of the major contributions of this research is to introduce and demonstrate by way of experimental results that techniques from the field of data mining are richer in extracting useful information and knowledge and constitutes a reliable and efficient carrier of the clinical decision support system. Thus, using data mining techniques new unobserved and unsuspected relationships and trends from previously existing data that can be used as input to develop strategies and policies of diarrheal disease control and prevention is extracted. Moreover, although the study is focused at addressing diarrheal disease

problems in particular, the output of the study may be used as a source of methodological approach for studies dealing with the application of data mining technology on similar problem areas.

1.6 Ethical Considerations

It is known that all fields/areas have their own ethics. Among these, the areas of health care are those with their own predefined ethical rules and regulations. These rules are concerned with the way of interaction between patients and health care professionals, managerial bodies, researchers, and other staffs with in and out of these organizations. Especially the inter communication and dissemination of information between the health care service providers and patients needs especial attention. Because, Privacy of records and ethical use of patient information by the health care professionals is always a major concern within these organizations. Likewise, researchers in related areas of health care may face these issues and challenges. But, it is always true that, researches are essential in all fields to alleviate problems raised and to find solutions which in turn promote the quality and effectiveness of services. For example, healthcare records are private information and yet, using these private records may help stop deadly diseases. The option is that, not using the core identifiers which may disclose the individual with his/her clinical information for other bodies. And it is obvious that one of the core identifiers of the individual is ‘name’. Thus, for the purpose of this research, the researcher did not include the name of patients when he was collecting data. Overall, the data collected from the concerned institution was obtained under a confidentiality agreement and remained in confidence.

1.7 Thesis Organization

This thesis is organized into five chapters. The first chapter deals with the general overview of the study including background, statement of the problem, objectives of the research. The second is devoted to literature review of data mining technology as well as diarrheal disease respectively. Chapter three explains the Research methodologies,

decision trees and Naïve Bayes classifiers as well as the Weka software, used in this study. Chapter four presents the experimentation phase of the study. It comprises training, building and validation of the models. Results of the experiment are also analyzed and interpreted. The last chapter is devoted for the final conclusions and recommendations based on the research findings.

CHAPTER TWO

LITRATURE REVIEW

2.1 DATA MINING

2.1.1 Introduction

In today's world, “knowledge is power”. It is well known that in an Information Technology (IT) driven society, knowledge is the most significant asset of any organization. This knowledge can be used for fast and better decision making (Wasan et al., 2006). It is a valuable asset to organizations as a substantial source to enhance understanding of data relationships and support useful decisions to increase organizational competency (Kerdprasop N. & Kerdprasop K.,2009).The data stored in these organizations is a strategic resource and is being an important source of knowledge. Making use of most of these strategic resources can lead to improve the effectiveness and efficiency of their services and products.

With the rapid computerization of businesses and organizations, a huge amount of data has been collected and stored, and the rate at which data stored is growing rapidly (Magendram, 2007). However, turning increasingly large amounts of data into useful insights and finding how to better utilize those insights in decision making remains a challenge for most (Ranjan et al., 2007). Because, traditional statistical techniques, methods and data management tools are no longer adequate for analyzing this vast collection of data. This lead to a need for technologies that can support a user by analyzing and transforming large amounts of data into useful information and knowledge (Hand et al., 2001).Thus, data mining (or knowledge discovery in databases (KDD) is emerged as a solution to address and alleviate such issues.

Data mining is “the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, and other information repositories”(Obenshain, 2004). It is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data (Hajizadeh et al., 2010). It aims to identify valid, novel, potentially useful, and understandable correlations and patterns in data by combing through copious data sets to sniff out patterns that are too subtle or complex for humans to detect (The 14th Pacific-Asia Conference, 2010). It is a term synonymous with data dredging or fishing and has been used to describe the process of trawling through data in the expectation of identifying patterns that can aid decision making (Untwal, 2008).

Data mining is the most important technology in the knowledge discovery process. It can be considered as the heart of KDD process (Eapen, 2004). It is also an emergent and rising area of research and development, both in academic world as well as in business, connecting interdisciplinary studies and development adjacent to diverse domains (Bresfelean, 2007). Typical problems that data mining addresses are how to classify data, cluster data, and find associations between data items (Silver et al., 2001).

Data mining has been used intensively and extensively by different organizations and institutions, for example, financial institutions, for credit scoring and fraud detection; marketers, for marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling; healthcare organizations, to identify factors that contribute to various diseases, to make customer relationship management decisions, identify effective treatments and best practices for customers and enabling them to receive better and more health care services (Kaur & Wasan, 2006).

2.1.2 Data Warehousing

A data warehouse is a central store and consolidated view of enterprises' data, and optimized for reporting and analysis. Different organizations set up a data warehouse when it is perceived that a body of data is critical to the successful running of their business. Such data may come from wide variety of sources, and is then commonly made available via coherent database mechanisms (Barker, 2011).

In data warehousing, data and information are extracted from heterogeneous production data sources as they are generated, or in periodic stages, making it simpler and more efficient to run queries over data that originally came from different sources. Data is turned into high-quality information to meet all enterprise reporting requirements for all levels of users. Companies may use data warehousing to view day-to-day operations, to analyze trends over time and to facilitate strategic planning resulting from long-term data overviews. From such overviews, business models, forecasts, and other reports and projections can be made. The data that is going to be mined to get patterns and trends is first extracted from an enterprise data warehouse (TCC, 2005).

2.1.3 Knowledge discovery in databases (KDD)

KDD has become a widespread activity undertaken by an increasing number and variety of industrial, governmental, and research organizations. Knowledge discovery is the process of identifying and finding valid, novel and understandable patterns and relationships in data (usually in large databases). Knowledge discovery from large databases, often called data mining, refers to the application of the discovery process on large databases or datasets (Glasgow & Ng., 1999). KDD is the process of finding useful information, where as data mining is the process for extracting knowledge (information and patterns) derived by the KDD process using algorithms (Wasan et al., 2006).

Data mining and knowledge discovery in databases relate to the process of extracting valid, previously unknown and potentially useful patterns and information from raw data

in large databases.“The analogy of “mining” suggests the sifting through of large amounts data to find something valuable. It is a multi- step, iterative inductive process (Gerver & Barrett, 2006). It includes such tasks as problem analysis, data extraction, data preparation and cleaning, data reduction, rule development, output analysis and review. By and large, data mining and knowledge discovery in databases are treated as synonyms and refer to the whole process in moving from data to knowledge (Kraft et al., 2002). It can help organizations turn their data into information and knowledge. The advancement of Software and hardware enable organization to tap the power of KDD using computers (DeGruy, 2000).

According to Glasgow and Ng. (1999), the knowledge discovery process can be broken into several steps, including: Learning and understanding of the application domain; creating a target data set; data cleaning and preprocessing; data transformation; data mining to search for patterns of interest; interpreting and evaluating discovered patterns; and using discovered knowledge.

- Learning and understanding the application domain: includes relevant prior knowledge and the goals of the application.
- Creating a target dataset: refers to selecting a dataset or focusing on a subset of variables or data samples on which discovery is to be performed.
- Data cleaning and preprocessing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields.
- Data transformation: data are transformed or consolidated in to forms appropriate for mining by performing summary or aggregation operations.
- Data mining: is a process where intelligent methods are applied in order to extract data pattern. It includes searching for patterns of interest in a particular representational form or a set of such representations, including classification

rules or trees, regression, clustering, sequence modeling, dependency, and line analysis.

- Interpretation/Evaluation: this indicates interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users. The evolution process is also carried out to identify interesting patterns representing knowledge based on some interestingness measures.
- Using discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge. Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

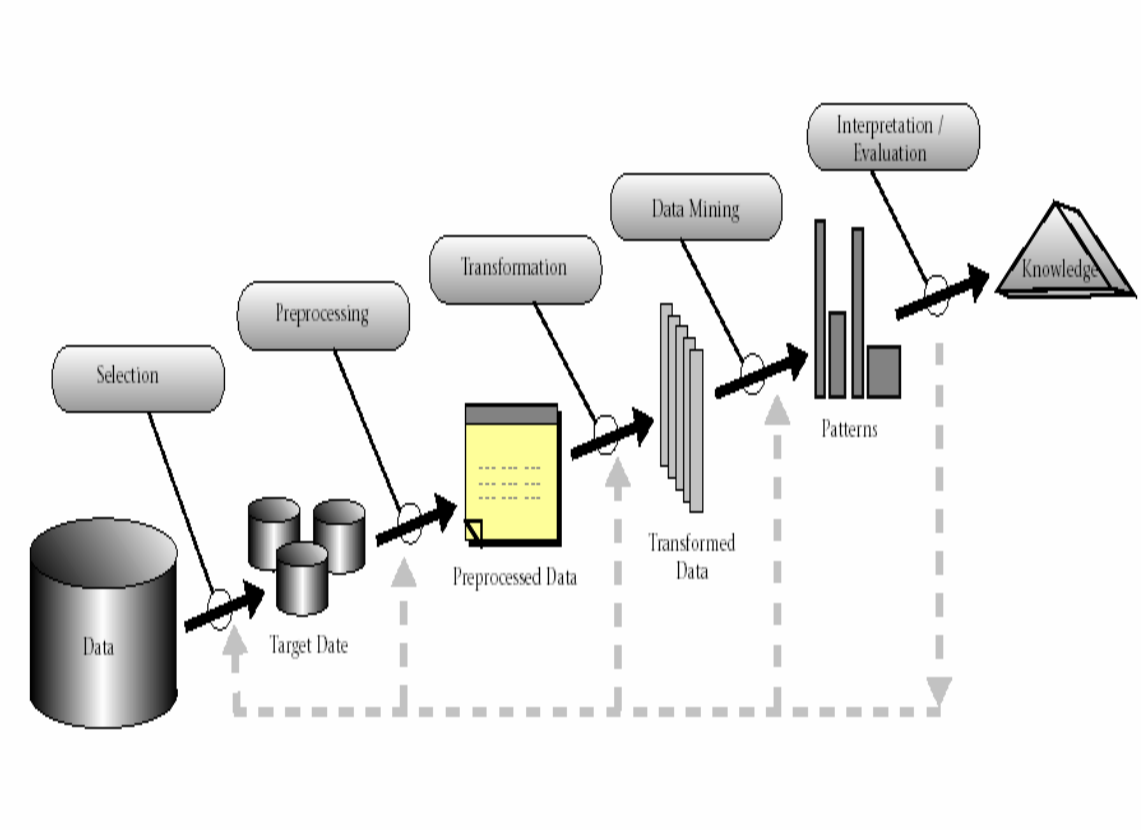

Figure 2.1: Overview of the steps involved in the KDD process (Eapen G., 2004).

2.1.4 Machine learning categories

Machine learning provides two important learning categories, namely supervised and unsupervised learning. These two learning categories are associated with different machine learning techniques that represent how the learning method works.

2.1.4.1 Supervised machine learning

With supervised learning there is a presence of the outcome variable to guide the learning process. There are varieties of supervised learning methods such as decision trees,Naïve Bayes, neural networks and KNN which attempt to discover the relationship between

the input variables and the class attribute . Given some training data described in terms of a set of features and their class labels, the goal of supervised learning is to find the partitioning of the attributes that allow correct classification of the training data as well as generalization from training data to unseen, similar data. Due to the presence of predefined examples or classes while learning, supervised learning is also referred to as classification (Roach & Maimon, 2005).

2.1.4.2 Unsupervised machine learning

Unlike supervised learning, unsupervised learning builds models from data without predefined classes or examples. This means, no "supervisor" is available and learning must rely on guidance obtained heuristically by the system examining different sample data or the environment .The output states are defined implicitly by the specific learning algorithm used and built in constraints (Caelli & Bischof, 1997)

Since the aim of this paper was to use classification techniques which fall under the supervised learning category, unsupervised learning was considered as beyond the scope of this thesis.

2.1.5 Data Mining Models

A model is a representation of the real world. Without a perfect and error free model or representation of the real world, well-intended decisions may lead to disastrous results. It forms the cornerstones of data mining. From a non-technical perspective, a data mining model may be considered as a black box, whose input is data to be mined and whose output is the knowledge discovered from the data. With this perspective, data mining employs two types of models: predictive and descriptive, which are like two sides of the same coin. One side, predictive models operate on certain attributes or characteristics of the data (independent variables) to predict other attributes (dependent variables). Descriptive models on the other side do the reverse, based on specific outcomes

(dependent variables), descriptive models attempt to look for the contributing attributes associated with the outcomes (dependent variables) (Glover et al., 2010).

2.1.5.1 Predictive Models

Predictive models are used to predict the value of a particular attribute. They perform inference on the current data in order to make the prediction. It is concerned with the construction of models that can be used to predict some object's target property from the description of this object. "Predictive modeling uses a set of tools to stratify a population according to its risk of nearly any outcome". For example, in the area of health care, it is used to predict a patient's susceptibility to a certain disease. That is, risk stratified and of patients to identify opportunities for intervention before the occurrence of adverse outcomes that result in increased deaths and medical costs (SIC, 2009).

Moreover, it helps to make a diagnosis of a particular disease. For example, a patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. In other areas, this model is used in predicting a long-distance customer's likelihood of switching to a competitor, an insurance claim's likelihood of being fraudulent, the likelihood someone will place a catalog order, and the revenue a customer will generate during the next year.

Often, Predictive models are used as descriptive. A predictive model's descriptive aspect is sometimes more important than its ability to predict. For example, suppose a researcher builds a model that predicts the likelihood of a particular cancer. The researcher might be more interested in examining the factors associated with that cancer or its absence than with using the model to predict if a new patient has the disease. Almost all predictive models can be used descriptively (Gerritsen, 1999).Tasks of predictive modeling includes: Classification, regression, and time series analysis.

2.1.5.1.1 Classification

Classification maps or classifies a data item into one of several predefined classes. It is a predictive modeling, in which we give a pre-defined grouping and try to predict the group of a new data point. A set of classification rules are generated from the classification model, based on the features of the data in the training set, which can be used to classify future data and develop a better understanding of each class in the database. For example, classification rules about diseases can be extracted from known cases and used for diagnosis of new patients based on their symptoms. This is the most important data mining technique, and medical diagnosis is an important application of classification. We may classify patients with heart problems on the basis of various types of heart diseases (Wasan et al., 2006).

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will act. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from a historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier (TCC, 2005).

Some knowledge of data under consideration is assumed before applying the classification technique. Suppose D is a database of patients. Someone may regard D as set of tuples (x_1, x_2 … x_n) where x_1, x_2 … x_n are values of attributes A_1, A_2 … A_n relevant to a particular disease. We may define various classes C= {C_1, C_2 … C_n} of patients depending on severity of disease or particular classification type of the disease (Wasan et al., 2006). Most of Classification and regression tools use Artificial Neural

Network, decision trees, Naïve Bayes, K-nearest neighbor, Regression and Time series techniques. For the purpose of this research, decision trees and Naïve Bayes are going to be more thoroughly explained in the next chapter.

2.1.5.1.1.1 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is an information processing paradigm inspired by information processing mechanisms of biological nervous systems (i.e., the brain). The key element of this paradigm is the structure of the information processing system modeled as a network comprising many highly interconnected Processing Elements (PEs) (also termed Neurons). Neurons work together in order to approximate a specific transformation function. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process during which the weights of the inputs in each neuron are updated (Elovici et al., 2007).

2.1.5.1.1.2 Decision trees

Decision tree learners are a well-established family of machine learning algorithms. Classifiers are represented as trees whose internal nodes are tests on individual features and leaves are classification decisions. The decision tree is induced from the dataset by splitting the variables based on the expected information gain. Modern implementations include pruning which avoids over-fitting. In this study J48, the Weka version of the commonly used C4.5 algorithm was evaluated. An important characteristic of Decision Trees is the explicit form of their knowledge which can be easily represented as a set of rules (Elovici et al., 2007).

2.1.5.1.1.3 Naïve Bayes

This technique limits its inputs to categorical data and applies only to classification. Named after Bayes's theorem, the technique acquired the modifier "naïve" because the algorithm assumes that variables are independent when they may not be. Simplicity and

speed make Naïve Bayes an ideal exploratory tool. The technique operates by deriving conditional probabilities from observed frequencies in the training data (Gerritsen, 1999).

2.1.5.1.1.4 K-nearest neighbor.

K-nearest neighbor (K-NN) technique differs from other techniques in that it has no distinct training phase; the data itself becomes the model. To make predictions for a new case using this technique, it is necessary to find the group with most similar cases ("k" refers to the number of items in this group) and use their predominant outcome for the predicted value (Gerritsen, 1999).

2.1.5.1.1.5 Regression

Regression is a method to map target data using some known type of function. It deals with estimation of an output value based on input values (Wasan et al., 2006).Regression uses existing values to forecast what other values will be. This method can be used to define the boundary condition by evaluating the data and determining the boundary through mathematical analysis (Trybula, 1997). In the simplest case, regression uses standard statistical techniques such as linear regression. The linear regression method is used for modeling continuous response. Unfortunately, many real world problems are not simply linear projections of previous values. Therefore, more complex techniques, such as logistic regression, decision trees, or neural nets, may be necessary to forecast future values (TCC, 1999).

2.1.5.1.1.6 Time series analysis

Time series analysis is the value of an attribute examined over a time period usually at evenly spaced time intervals. For example, depending upon the conditions of a patient, values of certain attributes may be obtained on a daily or hourly basis. This may be used to predict future values or to determine similarity between different time intervals (Wasan et al., 2006).

2.1.4.2 Descriptive models

Descriptive models describe or summarize the general characteristics or behavior of the data in the database. They are used to identify patterns in data. Descriptive modeling methods are also known as exploratory analysis or unsupervised classification; because there is no a prior knowledge available about the class labels to create the data model. They provide a summary or a human interpretable view of complex dataset and hence allow us to make decisions based on the data (Wasan et al., 2006).

Descriptive models encompass two important data mining tasks: These are clustering and association (Gerritsen, 1999).

2.1.4.2.1 Clustering

Clustering is also referred to as segmentation and it lumps together similar people, things, or events into groups called clusters (Gerritsen, 1999). It is the identification of classes or clusters for a set of unclassified objects based on their attributes. It is a knowledge discovery process to find groups of interrelated cases and the statistical behaviors that make them adhere into groups. It requires using a distance measure, like the nearest neighbor technique. Once the clusters are decided, the objects are labeled with their corresponding clusters, and common features of the objects in a cluster are summarized to form the class description. For example, a set of new diseases can be grouped into several categories based on the similarities in their symptoms, and the common symptoms of the diseases in a category can be used to describe that group of diseases (Wasan et al., 2006). Clustering requires significant involvement from a business or domain expert who must judge whether the resulting clusters are useful or not (Gerritsen, 1999).

2.1.4.2.2 Association

Association is the discovery of relationships or correlations between items in a database (Kerdprasop, N. & Kerdprasop, K., 2009). In recent years, mining association rules on large data sets has received considerable attention. An association rule is in the form of "A_1 ^A_2 …^ A_i => B_1 ^ B_2^…^ B_j" which means objects B_1 ^ B_2^…. ^ B_j (B_1 and B_2 …and B_j) tend to appear with objects A_1 ^ A_2 …^ A_i (A_1 and A_2 … and A_i) in the target data. For example, one may discover that a set of symptoms often occur together with another set of symptoms (Wasan et al., 2006).In data mining, association rule is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules shows attributed value conditions that occur frequently together in a given dataset. These rules are useful for determining correlations between attributes of a relation and have applications in marketing, financial, and retail sectors. For example, optimized association rules are permitted to contain uninstantiated attributes and the problem is to determine instantiations such that either the support or confidence of the rule is maximized. For example, data are collected in supermarkets. Such 'market basket' data bases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers could use this data for adjusting store layouts, increase cross-selling or anticipate demand, structure promotions, and catalog design and to identify customer segments based on buying patterns (Hajizadeh et al., 2010). It helps to understand what products or services customers tend to purchase at the same time. It is often called market basket analysis and generates descriptive models that discover rules for drawing relationships between the purchase of one product and the purchase of one or more others (Gerritsen, 1999).

2.1.6 Data mining Applications

Many organizations have been used data mining intensively and extensively (Koh & Tan, 2005). Data mining helps organizations to identify valuable customers, predict future

behaviors, and enable them to make proactive, knowledge-driven decisions (Rygielski et al., 2002).These organizations have employed data mining in a number of areas including finance, to detect fraudulent patterns of clients, identify correlations between financial indicators; insurance to predict behavior patterns of risky clients, analyses claims, predict which customers are likely to buy policies; web analysis, to assess user browsing patterns, customer support; transportation, to analyze loading patterns, determine distribution schedules among outlets etc. (Glover et al., 2010).

Other data mining applications, for example, In the case of marketing, to identify buying behavior of clients, find associations between clients' demographic characteristics and predict clients purchase patterns. Retailers can keep detailed records of every shopping transaction through the use of store-branded credit cards and point-of-sale systems which in turn enables them to better understand their various customer segments.

In the area of banking, to utilize knowledge discovery for various applications, including: Fraud detection which is enormously costly. By analyzing past transactions that were later determined to be fraudulent, banks can identify patterns. In case of predictive life-cycle management, data mining helps banks predict each customer's lifetime value and to service each segment appropriately (Rygielski et al., 2002).

Data mining for health care is an emerging field where researchers from both academia and industry have recognized the potential of its impact on improved health care by discovering patterns and trends in large amounts of complex data generated by health care transactions. For example, data mining can help to predict the likelihood of patients getting a certain disease and disease prevalence, in detecting disease diagnosis etc. (Androwich et al., 2003). It is used to characterize patient behavior and predict office visits, identify successful medical therapies for different illnesses. Data mining techniques enable decision makers to identify patterns in clinical, claims and activity based historical data, to better identify and understand explanatory relationships between

variables that describe operational processes. This information can be utilized to devise strategic initiatives to improve organizational performance (Kudyba & Gregorio, 2010).

Application of data mining technology in manufacturing has also enormous benefits. For example, through choice boards, manufacturers can customize products for customers; therefore they can be able to predict which features should be bundled to meet customer demand (Rygielski et al., 2002).

2.1.7 Data Mining Applications for Health Care

Like any other organizations, health care industries are collecting and storing data in their data to day activities. Over time, this data is becoming huge and difficult to analyze and get useful information using traditional methods. It has been estimated that an acute care hospital may generate huge data which close to terabytes a year and this grow year after year. The data may contain both patients' medical and demographic information and data from laboratories. In view of the large amount of medical data being generated, there is growing pressure for improved methods of data analysis and knowledge discovery (Wasan et al., 2006).Thus, advanced analytical technologies and tools are becoming highly applicable in this industry. Among these, the application of data mining technology for health care industry is now being popular. Data mining presents the techniques and tools used in converting large healthcare data into useful information for decision-making (Chye & Gerald, 2007).

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain (Wasan et al., 2006).It is an emerging area of computational intelligence applied to automatically analyze patients' records aiming at the discovery of new knowledge potentially useful for medical decision making. Induced knowledge is anticipated not only to increase accurate diagnosis and successful disease treatment, but also to enhance safety by reducing treatment-related errors (Kerdprasop, N. & Kerdprasop, K., 2009).

Medical data mining has been instrumental in detecting fraud and abuse. In this aspect, it is used to identify unusual patterns and uncover fraud from the mass of data generated by millions of prescriptions, operations and treatment courses. Among other things, these applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims (Koh &Tan, 2005). It has also aided in diagnosis of new diseases. For example, It is known that "Discovery of HIV infection and Hepatitis type C were inspired by analysis of clinical courses unexpected by experts on immunology and herpetology, respectively"(Eapen,2004).Moreover, it assists medical treatment effectiveness by comparing and contrasting causes, symptoms, courses of treatment and deliver an analysis of which courses of action prove effective (Cabell,2006). For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. Other data mining applications related to treatments include associating the various side-effects of treatment, pooling common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub-populations that respond differently from the mainstream population to certain drugs, and determining proactive steps that can reduce the risk of affliction (Baylis,2008).

Data mining applications can also be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. It enables to stratify patient populations by demographic characteristics and medical conditions to determine which groups use the most resources, enabling it to develop programs to help educate these populations and prevent or manage their conditions (Kerdprasop, N. & Kerdprasop, K., 2009).

2.2 DIARRHEA DISEASE

2.2.1 Introduction

Technically speaking, diarrhea is the presence of three or more stools passed in 24 hours that are sufficiently liquid to take the shape of the container in which they are placed (Keusch et al., 1993). According to WHO (2005), it is the passage of unusually loose or watery stools, usually at least three times in a 24 hour period. Frequent passing of formed stools is not diarrhea. Babies fed only breast milk often pass loose, "pasty" stools; this also is not diarrhea.

Diarrhea is caused by infectious organisms, including viruses, bacteria, protozoa, and helminthes that are transmitted from the stool of one individual to the mouth of another termed fecal-oral transmission (Keusch et al., 1993). Repeated attacks of diarrhea are dangerous and will lead to malnutrition and growth retardation due to food restriction, anorexia, and malabsorption (Rañoa, 1983). Diarrhea is dangerous because, there is a loss of more than usual amount of water and salts from the body which results in dehydration. It occurs when the output of water and salts is greater than the input. The more diarrhea stools a child passes the more water and salts he/she loses. Dehydration during diarrhea is very serious and is a leading cause of child death around the world. The other reason that makes it dangerous is loss of nutrients. Nutrients are lost from the body during diarrhea. Diarrhea can cause malnutrition which becomes worse if a mother does not feed her child while the child has diarrhea (USAID, 2005).

Diarrheal illness is a problem worldwide, with regional variation in the prevalence of specific pathogens, the availability of means of diagnosis and treatment, and the degree of prevention achieved. Across the globe there are an estimated 1.8 billion episodes of childhood diarrhea annually, mostly in developing countries (A. Mekasha & A. Tesfahun, 2003). Diarrhea and Pneumonia are responsible for an estimated 40% of all child deaths around the world each year. Diarrhea is more prevalent in the developing world. It is

mainly due to the lack of safe drinking water, sanitation and hygiene, as well as poorer overall health and nutritional status (UNICEF & WHO, 2009). Globally, more than 125 million children under-five years of age live in households without access to an improved drinking-water source, and more than 280 million children under-five live in households without access to improved sanitation facilities. These factors together contribute to about 88% of deaths from diarrheal diseases or more than 1.5 million children under- five perish from diarrhea each year (Girma et al., 2008).

According to UNICEF and WHO, an estimated 2.5 billion people of all ages lack improved sanitation facilities, and nearly one billion people do not have access to safe drinking water. These unsanitary environments allow diarrhea-causing pathogens to spread more easily. Improving unsanitary environments alone, however, will not be enough as long as children continue to remain susceptible to the disease and are not effectively treated once it begins. Only 39% of children with diarrhea in developing countries receive the recommended treatment (UNICEF & WHO, 2009). Thus, It is the leading causes of morbidity and mortality in these countries, especially among children under the age of five (Girma et al., 2008). Evidence has also shown that children with poor health and nutritional status are more vulnerable to serious infections of diarrhea and suffer multiple episodes every year. Diarrhea seriously aggravates poor health and malnutrition in children and creating a deadly cycle (UNICEF & WHO, 2009).

Children with poor nutritional status and overall health, as well as those exposed to poor environmental conditions, are more susceptible to severe diarrhea and dehydration than healthy children. Children are also at greater risk than adults of life-threatening dehydration since water constitutes a greater proportion of children's bodyweight. Young children use more water over the course of a day given their higher metabolic rates, and their kidneys are less able to conserve water compared to older children and adults (UNICEF & WHO, 2009).

Studies conducted in Asia, Africa and Latin America, have shown that around 750 million children below 5 years of age suffer from acute diarrhea each year (Rañoa,1983). In Africa, a child experiences five episodes of diarrhea per year, and 800,000 children die each year from diarrhea related dehydration (Girma et al., 2008). Diarrhea is also a common manifestation of HIV infection in both adults and children (UNICEF& WHO, 2009).

2.2.2 Diarrheal Disease in Africa and Risk Factors

Diarrhea has been estimated to be responsible for 50% of all childhood illnesses in Africa. Approximately 40% of childhood deaths from diarrhea worldwide was occured in Sub-Saharan Africa by the year 2000, although only 19% of the world's population under the age of five years was lived in this region. A number of different social, political, and economic factors are present in Sub-Saharan Africa which contributes to the constant morbidity from acute and persistent diarrhea, as well as intermittent epidemics of cholera and dysentery common to this region of the world (Hamer et al., 1998).

Broadly recognized risk factors for diarrheal diseases include little or no access to safe water and sanitation, as well as poor hygiene and feces disposal practices (DT et al., 2006). Approximately 50 percent (300 million individuals) of the African population have no access to safe water, and 66 percent (400 million individuals) lack access to hygienic sanitation (WHO,2000).These and many other factors, such as poor housing and crowding, are intrinsically associated with poverty. Furthermore, poverty usually limits access to health care and restricts appropriate and balanced diets. Inequities in exposure and resistance add up to inequities in coverage of available preventive interventions, access to an appropriate health provider, and care, making poor children more likely to become sick than the better-off children (DT et al., 2006). A WHO report on global water supply provides worrisome figures of current and future scenarios for Africa. Of all the regions in the world, the African region was the only one showing a decline in the

proportion of the population that had access to sanitation between 1990 and the year 2000 (WHO, 2000).

2.2.3 Diarrhea Disease in Ethiopia

Health service records and community based surveys in the past years indicate that diarrheal diseases are major causes of morbidity and mortality in Ethiopia (ERCS, 2010). It is the fourth leading causes of child mortality in the country followed by Pneumonia, Neonatal conditions and Malaria respectively. According to the report of FMoH, About 472,000 Ethiopian children die each year before their fifth birthday of which 20% is attributed by diarrhea (FMoH, 2005). Its prevalence is mainly caused by access to safe water, sanitation and hygiene problems. Only 55% of the general population has access to safe water and that percentage drops to 35% for those in rural areas. This lack of access to safe water and adequate sanitation increases the morbidity and mortality from diarrheal disease (ERCS, 2010).

In 2006, Over 100,000 cases and 1,200 deaths were reported in the country (WHO, 2008). According to the report of WHO (2007), that comprised from august to October, 6 out of 9 regions and Addis Ababa City were affected by acute watery diarrhea (AWD). The affected regions were Oromiya, Southern Nations Nationalities and People's Region (SNNPR), Amhara, Afar, Tigray and Somali. 129 districts were Affected in the 6 regions and 6 sub-cities in Addis Ababa.There were 2,040 cases and 43 deaths reported as of 30 September 2007.There were 5,823 cases in August 2007.There were also a continuity of reporting by the aforementioned 6 regions and Addis Ababa City for AWD cases (WHO, 2007). A study carried out in 2008 also revealed that the two-week period prevalence of diarrhea in under-five children was about 30.6% and 17.7% in Ethiopia and Oromia region, respectively (Girma et al., 2008). According to the report of FMoH in 2009, acute watery diarrhea (AWD) cases had rapidly increased from July to August 2009 and a total of 11,667 cases with many deaths in these regions were reported from 35 districts in the

same regions except Tigray mentioned above. In the same year, FMoH projected that a number of cases may reach 130,000 affecting 90 Woredas if adequate response operations were not taken (ERCS, 2010).The outbreak of this disease in Addis Ababa in the same year resulted in 15 deaths from persons close to 1,000 as a result of contaminated drinking water (Desalegn, 2009).

2.2.4 Clinical Types of Diarrhea

It is most practical to base treatment of diarrhea on the clinical type of the illness (WHO, 2005). Accordingly, there are three major clinical syndromes/types of diarrhea. They are acute watery diarrhea, persistent diarrhea, and bloody diarrhea.

Acute watery diarrhea is often characterized as urgent, frequent and watery bowel movement. The condition usually lasts for less than 2 weeks. According to the Definition of WHO (2005), diarrhea acute in nature is called acute watery diarrhea. It is more prevalent in people who live in unhygienic conditions and are more exposed to contaminated water or foods.

Persistent diarrhea begins acutely and lasts at least 14 days. It is typically associated with malnutrition, either preceding or resulting from the illness itself. It is usually associated with weight loss and, often, with serious non-intestinal infections Persistent diarrhea occurs more often during an episode of bloody diarrhea than an episode of watery diarrhea, and the mortality rate when bloody diarrhea progresses to persistent diarrhea is 10 times greater than for bloody diarrhea without persistent diarrhea (Keusch et al.,1993).

Bloody diarrhea is often referred to as dysentery and is marked by visible blood in the stools. It is associated with intestinal damage and nutrient losses in an infected individual [UNICEF & WHO, 2009). Nutritional status has profound effects on the duration and severity of bloody/dysenteric syndromes (Sazawal et al., 2011).

These clinical types of diarrhea can be treated either in the home or health facilities depending on their severity. If their symptoms are not as such danger on the patient with diarrhea, they can be treated in the home, which is referred to as home therapy; otherwise, the patient should be taken to health facilities. In the home, for example, appropriate and suitable fluids, nutritious diet, vitamins and minerals can be used to prevent and reduce diarrheal disease risks. In health facilities, more scientific treatments can be given if and only if the home therapies could not alleviate and reduce the severities of the disease (WHO, 2005). Zinc and vitamin A supplementation, ORT/ORS and antibiotics are some of treatments given in health facilities (DT et al., 2006).

In general, a child with diarrhea should be assessed for type of diarrhea, dehydration, malnutrition and serious infections; so that an appropriate treatment can be developed and implemented without delay. Understanding the history of the child and physical examination of the child are used to the assessment. The history of a child assists to know: pre-illness feeding practices, duration of diarrhea, presence of blood in the stool, number of watery stools per day, presence of fever, cough, or other important problems (WHO, 2005).

2.2.5 Dehydration

During diarrhea there is an increased loss of water and electrolytes (sodium, chloride, potassium, and bicarbonate) in the liquid stool. Dehydration occurs when these losses are not replaced adequately and a deficit of water and electrolytes develops quickly (WHO, 2005). Unless these lost fluid and electrolytes replaced promptly, the body cannot function properly. Dehydration is particularly dangerous for children, who can die from it within a matter of days (NIDDK, 2003).

Dehydration can be graded as: No, Some and severe dehydration. Understanding the degree of dehydration is used to treat or prevent dehydration of patients with diarrhea. The degree of dehydration is graded according to signs and symptoms that reflect the

amount of fluid lost. In the early stages of dehydration, there are no signs or symptoms and thus, if there are No enough signs to classify as some or severe dehydration it is graded as no dehydration. For some dehydration, dehydration increases, signs and symptoms develop and the child may indicate 2 or more of the following signs: restless/irritable, sunken eyes, drinks eagerly/thirsty, when pinched, skin goes back slowly (1 second). In severe dehydration, effects observed in some dehydration become more pronounced and the patient may develop signs like being very sleepy and diminished consciousness, lack of urine output, cool moist extremities, low or undetectable blood pressure, sunken eyes, Not able to drink or drinking very poorly, When skin pinch goes back to normal very slowly (longer than 2 seconds (USAID, 2005).

2.2.6 Diarrhea and Malnutrition

Diarrhea is, in reality, as much a nutritional disease as one of fluid and electrolyte loss. Children who die from diarrhea, despite good management of dehydration, are usually malnourished and often severely so. During diarrhea, decreased food intake, decreased nutrient absorption, and increased nutrient requirements often combine to cause weight loss and failure to grow. The child's nutritional status declines and any pre-existing malnutrition is made worse. In turn, malnutrition contributes to diarrhea which is more severe, prolonged, and possibly more frequent in malnourished children (WHO, 2005). Evidence from numerous studies of children under five years of age in developing countries suggests that both acute and persistent episodes of diarrhea predispose to or exacerbate malnutrition, and conversely chronic malnutrition may be a risk factor for diarrhea (H. Hamer et al.,1998).That is, diarrhea is a serious and often fatal event in children with severe malnutrition. Malnutrition can be prevented and the risk of death from a future episode of diarrhea is much reduced by continuing to give nutrient rich foods during and after diarrhea, and giving a nutritious diet, appropriate for the child's age. Although treatment and prevention of dehydration are essential, care of these

children must also focus on careful management of their malnutrition and treatment of other infections (WHO, 2005).

2.2.7 Interventions to Control Diarrheal Diseases

Interventions are used in preventing and reducing diarrhea disease prevalence and outbreak that helps to make children and other age groups healthier and less likely to develop infections that lead to diarrhea (UNICEF & WHO, 2009). There is sufficient evidence that home therapy interventions are effective in the prevention and control of diarrheal diseases. These interventions include: exclusive breastfeeding, safe water, good feeding practice, good personal and community sanitation and hygiene (DT et al., 2006).

Breast feeding is important in preventing and controlling the occurrence and severity diarrhea among children who are breast fed. Especially during the first 6 months of life, it necessary to give exclusive breastfed for infants. Exclusively breastfed babies are much less likely to get diarrhea or to die from it than babies who are not breastfed or are partially breastfed. Breastfeeding should continue until at least 2 years of age (UNICEF & WHO, 2009).

The risk of diarrhea can be reduced by using the cleanest available water and protecting it from contamination (WHO, 2005). Interventions to improve water quality at the source, along with treatment of household water and safe storage systems, have been shown to reduce diarrhea incidence by as much as 47%.

Good feeding practices involve selecting nutritious foods and using hygienic practices when preparing them. The choice of complementary foods will depend on local patterns of diet and agriculture, as well as on existing beliefs and practices. In addition to breast milk (or animal milk), soft mashed foods (e.g. cereals) should be given. If possible, eggs, meat, fish and fruit can be also given. Other foods, such as well cooked pulses and vegetables are also important (UNICEF & WHO, 2009).

Personal and community sanitation and hygiene practices are also used to prevent diarrheal disease prevalence and outbreak. Good Personal and community sanitation and hygiene practices include: hand washing, food safety, use of latrines etc. (WHO, 2005).

2.3 Related Researches

Even if the researcher could not find specific studies which have used data mining technology in exploring diarrheal disease prevalence from the clinical data stored in previous years, many epidemiological researches related to diarrhea disease using other research methods have been done. Moreover, the researcher found and reviewed various health related researches done by applying data mining techniques. In the following section, some of the studies that have been done by different researchers using both of the above methods are discussed.

Stolba and Tjoa (2005) studied on the role of data warehousing and data mining technique for the use of evidence-based medicine. According to them, treatment without evidence may elongate the process of treating patients which in turn create a problem of getting necessary interventions on time .They added that, only external evidence-based knowledge is not enough for efficient treatment of individual patients. The support of Information Technology (IT) in the process of health care is underlined by these researchers. That is, development of evidence-based guidelines, support of the clinicians at the point of care and controlling of clinical pathways are undertakings which cannot be fulfilled satisfactorily without IT. Therefore, they proposed that, data mining, which is one of the applications of IT, is a very important and suitable tool to get necessary knowledge from immense data volumes in health care.

Eapen (2004) on his master's thesis entitled "Application of Data mining in Medical Applications"described that, even though healthcare data is a good test bed for data mining ,huge amount of data in the health care industry is still being collected and organized using hard copy materials, that is, using paper and pen, which are difficult to data mining process. He was using different data mining techniques in his paper and found that decision tree experiments were the most useful and informative experiments .

Palaniappan and Awan (2008) on their study of “Intelligent Heart Disease Prediction System Using Data Mining Techniques” found that, Naïve Bayes is the most effective technique followed by Neural Network and Decision Trees to predict patients with heart disease. Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors. They had five mining goals of which four are answered by Naïve Bayes, three and two are answered by Decision Trees and Neural Network respectively. Relationship between attributes produced by Neural Network is more difficult to understand. However, all the three techniques are able to extract patterns in response to the predictable state. Besides, he pointed out that decisionTrees results are easier to read and interpret .

Patil and Kumaraswamy (2009) used data mining technology to predict the different risk levels of heart attack. They used neural network techniques with Back-propagation and K-Means as training algorithm. Experimental results using the above techniques have illustrated the efficacy of the designed prediction system in predicting the heart attack. These researchers also concluded that health care management using data mining technology is not as easy as other fields due to the reason that the data existing here are heterogeneous in nature and that a set of ethical, legal, and social limitations apply to private medical information.

Felkey et al.(2003) conducted a study on data Mining for the Health System Pharmacist and they described that pharmacists can able to detect hidden knowledge buried in a body of clinical, pharmacy, and administrative data that is too large for human beings to investigate without a computer’s assistance. According to these researchers, decision-making abilities can be strengthened and the quality of pharmaceutical care they provide can be improved tremendously by using data mining technology. They added that pharmacists working in health care related areas should be ready to take up and apply data mining tools and techniques to improve their every day practice.

Bellaachia and Guven (2005) studied on predicting breast cancer survivability using data mining techniques. They used three data mining algorithms, such as, Naïve Bayes, neural net and C4.5 and compare the predicting ability of each algorithm. Their goal was to have high accuracy, besides high precision and recall metrics. They used these metrics related to other metrics such as specificity and sensitivity which can be derived from the confusion matrix. Even if neural net and C45 have comparable performances; they found out that C4.5 algorithm has a much better performance than the other two techniques. They also pointed out that neural net takes more computation times as compared to the other two. It takes 12 hours; where as Naïve Bayes and C4.5 took 1 minute and 1 hour respectively.

Srinivas et al. (2010) studied on data mining applications in health care in general and in prediction of heart attacks in particular. They have found that data mining techniques such as Rule based Decision tree, Naïve Bayes and Artificial Neural Network have a potential use in predicting and classifying of health related problems from massive volume of healthcare data.

Shegaw (2002) applied data mining technology to predict child mortality patterns up on community-based epidemiological datasets collected by the Butajira Rural Health Project (BRHP) epidemiological study. The methodology which was used by the researcher had three basic steps. These were collecting of data, data preparation and model building and testing. BrainMaker and See5 softwares were employed by the researcher so as to build models using neural net work and decision tree techniques respectively. He found that both the techniques yield comparable results for misclassification rates. However, unlike the neural network models, the results obtained from decision tree models provided simple and understandable rules that can be used by any health care professionals to identify cases for which the rule is applicable. In fact, the accuracy obtained from decision tree models also outperforms neural networks. He also found that best models using neural network technique by modifying default parameters of the program.

Moreover, he proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities.

A cross-sectional survey by G.Mitikie (2001) indicated that the use of unprotected water sources was significantly associated with diarrheal morbidity. In this survey, a high prevalence of dysentery/bloody and persistent diarrhea were observed.

Another cross-sectional Community Based Study by Mekasha and Tesfahun (2003) revealed that well source of water, lack of complete immunization, attack of measles and acute respiratory infections (ARI) were to be significantly associated with occurrence of diarrheal disease and persistent diarrhea was very high.

Yassin (2000) studied on morbidity and risk factors of diarrheal diseases among under five children used a community based cluster survey. He revealed that variables like child's age, mother's and father's education status and access to safe water were some of the factors appeared to be significantly associated with diarrhea prevalence. He also described that infants were found to have 4.6% higher risk of recurrent diarrhea than older children.

Andualem and abera (2010) conducted a community based descriptive cross-sectional study on the impact of latrine utilization on diarrheal diseases. They revealed that latrine utilization play a critical role in preventing and controlling diarrheal disease prevalence.

As an intervention for diarrheal disease, (Damte et al., 2008), conducted a case control study and described that Zinc supplementation in the treatment of childhood diarrhea is well tolerated by patients. A cross sectional study conducted by (Mirgissa & Fikau A, 2000), revealed also ORS as one of the major intervention mechanism for diarrhea.

Although the application of data mining technology in research area of many health related issues is rising, to the best knowledge of the researcher, studies related to

diarrheal disease using data mining technology were not carried out before. The studies that the researcher reviewed related to this disease especially in Ethiopia are simple surveys using traditional statistical techniques. That is, the huge data collected on diarrheal disease patients was not used to get useful information and knowledge using data mining technology. Thus, it is necessary to apply data mining technology to extract useful information and knowledge from the previous stored data related to diarrheal disease, and it is why, the researcher is doing this study concerning this issue.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

Today, the application of data Mining technology in research areas is rapidly increasing in various sectors. Health care industries are among those sectors which are using data mining technology extensively. Because, issues related to health care are increasing the amount of data stored in these sectors. Thus, the application of data mining technology as a research area for these sectors is very important. Data mining technology as a tool has its own methods, procedures and techniques to be followed and used in research. These methods, procedures and techniques may be chosen as per the nature of the data and the objectives of the researcher to be used for a specific study. In this research, the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology is used. It is the most commonly used methodology for developing data mining research projects. This model describes the activities that must be done to develop a data mining research projects. CRISP-DM has objectives such as ensuring quality of data mining research project results; reducing skills required for data mining; capturing experience for reuse; being general purpose (i.e., widely stable across varying applications and robust (i.e., insensitive to changes in the environment); tool and technique independent and tool supportable. One important factor of CRISP-DM success is the fact that CRISP-DM is industry-tool and application neutral (Mariscal et al., 2010).

3.2 Cross-Industry Standard Process for Data Mining (CRISP-DM) Process Model

CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. It consists on a cycle that comprises these six stages. These phases help different sectors understand

the data mining process and provide a road map to follow while planning and carrying out a data mining project. This model encourages best practices and offers organizations the structure needed to realize better and faster results from data mining (Shearer, 2000).

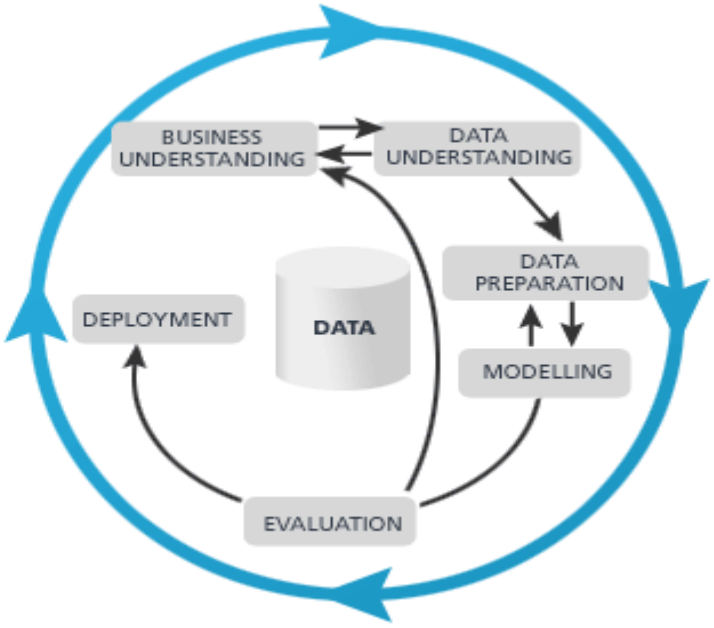

Figure 3.1: The CRISP-DM process model (Mariscal et al., 2010).

In the following sections, these phases are discussed; including the tasks involved with each phase in connecting with this study .The proposed method of this study was based on the classification and prediction task of data mining. Thus, Classification algorithms such as J48 decision tree and Naïve Bayes were those specifically used in this research. A decision tree is one of learning algorithms which poses certain advantages that make it suitable for discovering the classification rule for data mining applications. The naive Bayes classifier is also proved to be very effective on many real data applications. Moreover, operations of data mining tasks require specific softwares that contain these techniques. For the purpose of this research, the Weka software was used.

In general, this chapter elaborates the methods, procedures, techniques, software, and the nature and preparation of data which was used in this research.

3.2.1 Identifying Data Sources and Business Understanding

The data that was explored to conduct the study is from the patients' records of diarrheal disease control and training center of African sub Region II in Tikur Anbessa hospital. The center was under the supervision of the pediatrics department. The department was involved both in training of students and clinical services for children under Addis Ababa University. Because, the center is purposely built only for diarrheal cases, it helped the researcher to get representative data for the selected study site.

After the data sources were identified, it is necessary to understand the business of which the data was going to be collected. Business Understanding is the initial and most important phase of any data mining project. This phase focuses on understanding the data mining objectives from the business prospective (Shearer, 2000). Thus, before performing any analysis of datasets in the real world data in data mining, it is crucial to understand the need to data mining.

As mentioned in chapter 1, section 1.3, the main objective of this research project was to investigate the potential applicability of data mining techniques in exploring the prevalence of diarrheal disease. Classification and prediction techniques of data mining technology were used in extracting useful and interesting patterns and relationships among features of the dataset. Provided that useful relationship and rules among features were to be established, prevention programs for diarrheal disease could be established and have a better understanding of the nature and trend of the disease in Addis Ababa and in the country in general. This assists to develop strategic solutions to avoid and protect the vulnerable groups. After understanding the problem and the goal of the data mining task is defined, the researcher can easily select and understand the data that would be relevant for the intended purpose.

3.2.2 Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect appropriate subsets of the data being collected. Before starting the actual data mining task, we should be able to clearly define our problem and also have a good understanding of our data to be used for the data mining task.

The initial data collection involves selecting a representative section of the data which is likely suitable and reliable for the objectives stated. For this research, the data collection process was carried out initially by converting the paper based format data that was stored in the center in to a computer/electronic format (usually in a spread sheet).By the time of data collection, features or attributes and records which were irrelevant for the data mining process were excluded.

The data collected and used in this study included patients' records of diarrheal disease which was collected from the year 1995 to 2000. Each record of the data contained the following information: Year of treatment, Age, Sex, weight, Visits, Nutritional status, Type of diarrhea, Degree of dehydration and Treatment.

3.2.3 Data Preparation

One of the most important tasks in doing research related to data mining is preparing the data in a way that is acceptable for the specified data mining tools, techniques and tasks. The purpose of this stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in subsequent stages. CRISP-DM defines certain tasks in this phase which are very specific to "structured"data stored in a database like select data, clean data, construct data, and integrate data. Thus, after the data was collected, the researcher prepared the data in such a way that the data was appropriate to the requirements of the selected data mining tasks and the specific data mining tool. At the

time of data preparation, the researcher inspected the relevance of individual attribute values and types, quantity and distribution of missing values and noisy data. After that, all constraints related to the collected data were avoided using different mechanisms as per the requirements of the selected techniques. Thus, data preprocessing is the main task that was performed to alleviate the aforementioned data set constraints.

3.2.3.1 Data preprocessing

All raw data sets which are initially prepared for data mining are often large; many are related to humans and have the potential for being messy. Real-world databases are subject to noise, missing, and inconsistency due to their typically huge size, often several gigabytes or more. If there is much missing, irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Moreover, data mining tools may not accept dataset which is not error free. It is well known that data preparation and filtering steps take considerable amount of processing time in data mining tasks.

Data preprocessing is commonly used as a preliminary data mining practice to overcome these constraints. It transforms the data into a format that will be easily and effectively processed by the data mining software which in turn understandable for the users. There are a number of data preprocessing techniques which include: Data cleaning; that can be applied to remove noise and correct inconsistencies, outliers and missing values. Data integration; merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations; such as normalization which can improve the accuracy and efficiency of mining algorithms. Data reduction; can reduce the data size by aggregating, eliminating redundant features. The data processing techniques, when applied prior to mining, can significantly improve the overall data mining results.

The data collected in the diarrheal disease control and training center, which was employed for the purpose of this study, suffers from different constraints. These include missing values and encoding inconsistency in various attribute values.

These constraints resulted in difficulties to perform and fulfill predefined data mining objectives and tasks. Thus, to construct an optimal model, clean and automated data should be prepared. Thus, to achieve the best performance for a selected data set, pre-processing is an important process in data mining tasks (Nasereddin, 2009).

Therefore, the researcher of this study was used necessary data preprocessing techniques as needed depending on the performance of the data for the mining process. For example, various missing value techniques can be used for handling missing data for existing databases and for data left unknown during or not applicable during entry (Pedarla,2004).These include:

- Ignoring the instance: The record containing the missing value attribute is ignored/ omitted. This results in loss of a lot of information.
- Manual Replacement: Manually search for all missing values and replace them with appropriate values. Mostly, these are done when the replacing missing values are known.
- Using a global constant: Replacing all missing values with some constant like “unknown”or “?”.
- Using attribute mean/mode: Replacing the missing values with mean or mode of non missing values of that attribute or of same class.
- Using the most probable value: Replacing by the most probable value, using decision trees or Bayesian methods.
- Expectation maximization (EM) method: It proceeds in two steps. First step compute the expected value of the complete data record likelihood and the second step, substitute the missing values by the expected values, obtained from

the first step, and maximize the likelihood function. These steps are continued until convergence is obtained.

In the dataset collected for this study, there were attributes with missing values. To handle these missing values for the purpose of this study, attribute mean/mode and the most probable value techniques were used for numerical and non-numerical values respectively. Replacing the missing values with the mean or the mode of that attribute is the most common method of filling the attributes efficiently without too much computation. The Weka software includes a predesigned algorithm, Replace Missing Values Filter class, for handling the missing values. It uses the mean/mode method to handle the missing values, where the numeric missing values are replaced with their means and nominal/categorical values are replaced with their mode values.

The other task that was performed to make the data more suitable and understandable to the tool employed and to the data miner was attribute value derivation. The attribute 'Nutritional status' was originally recorded in interval form to assess the effect of diarrhea on patients' nutritional status when they came to the center for treatment. The attribute was placed as '<60, 60-80 and >80' on the paper file when this dataset was collected. By consulting domain experts, these intervals were replaced as 'Low, Medium and High' respectively.

There was also data value inconsistency in some attributes of the original data. For instance, to refer the 'age' of patients with diarrhea, day, month and year were used. Thus, this form of data placement is not suitable for data mining task. It was necessary to make this form of data in to one measurement. Since most of the ages of patients were placed in 'Month', it was better to transform the other measurements of age in to month for uniformity. Hence, all values of the 'age' attribute were changed in to 'Month' for this study.

The other task that was performed in preprocessing stage was class label balancing using the ‘SMOTE’ method from the preprocess package in Weka. This was done so as to resample and balance the number of class labels for those attributes used as target class in model building process. The target classes were ‘Treatment’ and ‘Type of diarrhea’. Both of them had three class labels with imbalanced size which directed the model to predict for those high in number. Before using the ‘SMOTE’ method, the class labels ‘ORS’, ‘ORS and RL’ and ‘RL’ for ‘Treatment’ attribute class had the size of 4334,158 and 1080 respectively. The class lables ‘Watery’,’Bloody’ and ‘Persistent’ of the ‘Type of diarrhea’ attribute class had also the size 5036,358 and 178 respectively. After the ‘SMOTE’ method was used, the size of class labels for ‘Treatment’ and ‘Type of diarrhea’ class attributes were balanced to 4334, 5056, 4320 and 5036, 5728 and 5696 in the order mentioned above respectively. Figures depicted below show the distribution of class labels before and after ‘SMOTE ‘was used for each target class.

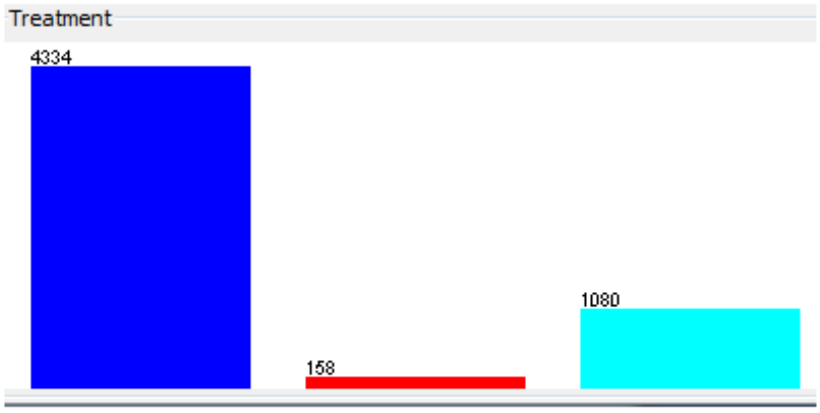

Figure 3.2: Statistic about class labels distribution in a data set based on ‘Treatment’ as a target class before ‘SMOTE’ was used.

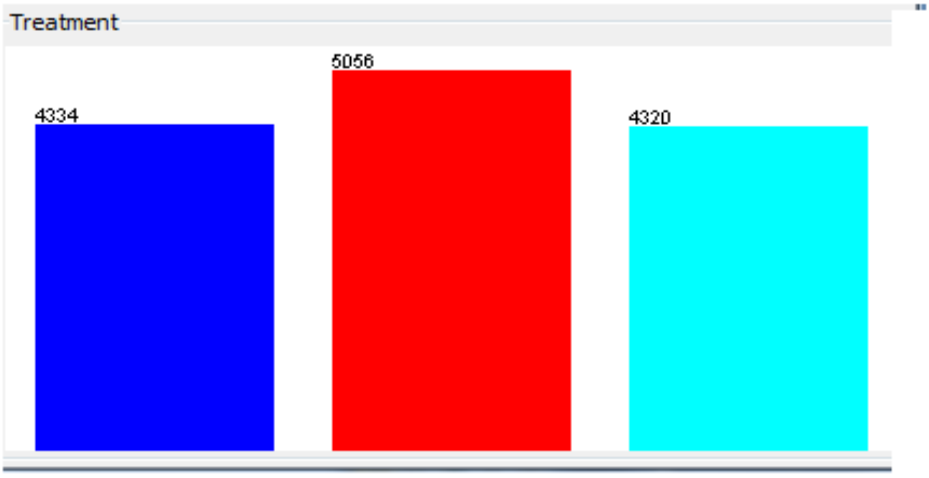

Figure 3.3: Statistic about class labels distribution in a data set based on 'Treatment' as a target class after 'SMOTE' was used.

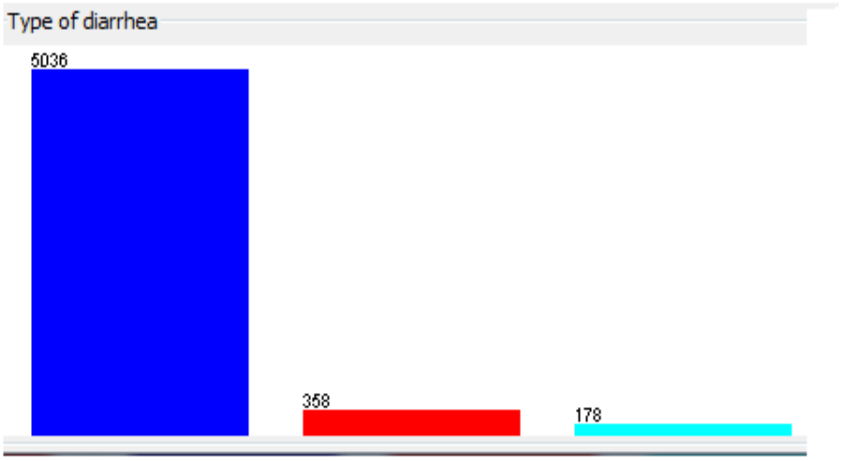

Figure 3. 4: Statistic about class labels distribution in a data set based on 'Type of diarrhea' as a target class before 'SMOTE' was used.

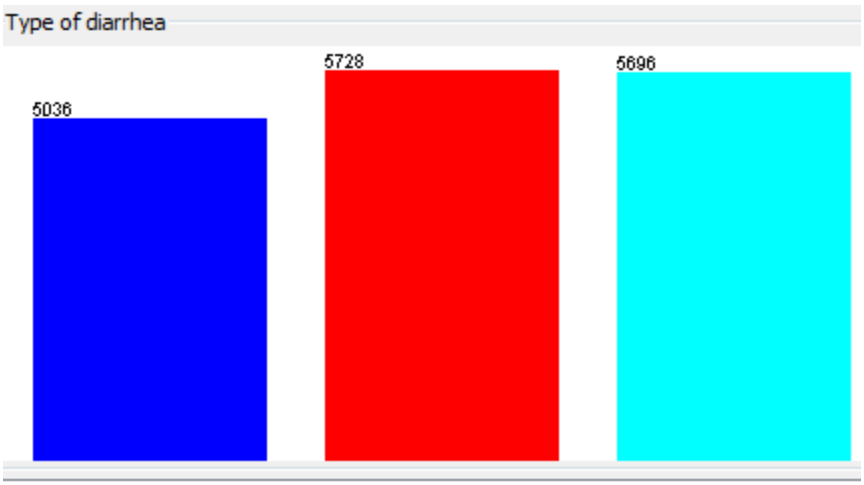

Figure 3.5: Statistic about class labels distribution in a data set based on 'Type of diarrhea' as a target class after 'SMOTE' was used.

3.2.3.1.1 Data Selection

Because of different factors, it was not possible to use the entire data already stored in the targeted data sources for data mining purposes. Insufficient time and budget as well as data quality by itself might be reasons for data selection. As explained in the above section, the data employed in this research is collected from the diarrheal disease control and training center of African Sub Region II in Tikur Anbessa Hospital. The center was established twenty years ago. Thus, clinical data related to diarrhea was stored for at least twenty years in the center. However, due to the constraints mentioned above as well as lack of organized data before the year 1995, the data employed for this study was only from 1995-2000 E.C. The other factor that restricted the researcher from using data collected in recent years (after 2000) was that the paper based files after this year are just in use by the workers to register day to day clinical data.

3.2.3.1.2 Attribute Selection

Removing unwanted attributes, irrelevant for the research goal should be considered in the construction of the final data set**.** Theoretically, some classifiers such as decision trees

determine relevant attributes for classification automatically using the concept of information gain or entropy without manual efforts. However, it is important to exclude those attributes that are not relevant for analysis in order to simplify the tasks performed before model building is started.

Although the number of attributes employed in this study was small, it was still necessary to select attributes which were relevant and important for the study. To decide on the relevant attributes for this study, a discussion with domain experts and reviewing of various materials was made. For example, one of the attributes that were ignored was 'Address'. Because, the values in the 'Address' attribute were not recorded consistently. Some of the values for this attribute identify the patient using Wereda and Kebele which was recorded numerically. The rest of the values identify patients using Sub City (Kifle Ketema) with alphanumeric data type. Therefore, based on the domain experts' opinion and the researcher's own observation, it was impossible to use these attribute in the model even if the attribute by itself was important to the classification and prediction purpose.

The other irrelevant attributes which were not collected include name, identification number (ID No) of patients and specific date when treatment was given for each patient. This attributes were excluded because of their irrelevancy for the study's objectives. For example, the patients' 'ID No' was not important and did not have any value in the output/results of mining process. That is, it did not have any role in relation with other variables in classification and prediction of clinical records of diarrheal disease. Likewise, the specific date when the treatment is given and patients' name did not have any contribution for the intended study and they were excluded from the dataset. In fact, Ethical considerations also restricted from using individual identifiers or sensitive personal references like 'Name'. The selected attributes for model building with their data type and description are depicted in appendix part.

Moreover, attribute value inconsistency was faced with in one of the critical attributes, that is, 'Type of diarrhea'. As a category for the classes of this attribute, in few of the pages of the paper file; Acute, Bloody and Persistent were used. In most of the other pages, the attribute was classified as Watery, Bloody and Persistent. Therefore, it was not only enough to discuss with the domain experts, but also reviewing of various materials was important to confirm and decide on these attribute values. Accordingly, the researcher got confirmation from both the domain experts and the materials reviewed which ensures, as discussed in chapter two, the attribute values are classified as the second category listed above. Under this variable, one value labled as 'Mucoid' was also detected as an outlier and removed from the dataset. Because, it was located only in one record out of the whole data collected.

3.2.4 Training and Building Models

One of the important tasks to be performed at this step is selection of relevant software that supports the data mining techniques which are to be employed in the study. In conducting this research Weka software version 3.6 was employed for reasons of accessibility and familiarity. This software package encompasses different techniques and algorithms. However, in this study only two techniques were used. One was J48 which is a decision tree classifier and used for decision tree construction. The other was Naive Bayes, where the estimation of the likelihood is performed by means of the simplistic (naive) assumption that an attributes is independent of each other, given the class. The detailed description of Weka software and the above stated classifiers is presented in this chapter, section 3. 3.1, 3.3.2 and 3.3.3.

3.2.5 Evaluation

At this stage, models which appear to have high quality are built. 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 sure it properly achieves the

business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached (Mariscal G. et al., 2010).Thus, for the purpose of this study, various models were built and evaluated. Finally, the models which are considered best by the researcher were used for further stages of the study. The evaluation was made based on selected outputs provided by models from each classifier.

3.2.6 Deployment

The creation of the model is not the end of the project. Even if the purpose of the model is to increase knowledge of the data, it will be necessary to organize the knowledge extracted, as well as to present it in a useful way to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it is the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models (Mariscal G. et al., 2010). For the purpose of this research, results of the study are reported and recommendations based on the findings of the study are given.

3.3 Data Mining Tool Selection

To perform any data mining task, selection of appropriate tools and techniques is an important task which may be initiated after the definition of problem to be solved and the related data mining goals. Various types of tools may be used for data mining task by different researchers. Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. Moreover, the accessibility and familiarity of these tools for the researchers can be also another factor for selection. Depending on these factors, Weka and Microsoft- Excel were very effective tools for conducting this research. The Weka software comprises different techniques and algorithms that should

be selected as per the objectives of the researcher. Accordingly, as explained in the above section, J48 decision tree and Naive Bayes algorithms/classifiers were selected for the purpose of this research. More details of Weka software and these algorithms are presented in the following sections.

3.3.1 The Weka software

Weka software is one of data mining softwares that are used for data mining research purposes. The Weka software is developed by researchers at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment for Knowledge Analysis. The system is written in Java which is widely available for all major computer platforms (Witten et al., 2000). It provides extensive support for the whole process to implement data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing both the input data and the result of learning. Weka includes a variety of tools for preprocessing a dataset, such as attribute selection, attribute filtering and attribute transformation, feeding into a learning scheme, and analyze the resulting classifier and its performance. Weka is organized in packages that correspond to a directory hierarchy. It consists of four graphical user interface modules available to the user .These are: Explorer, Experimenter, Knowledge Flow and Simple Command-line interface.

The explorer of Weka interface is the main module for visualizing and preprocessing the input data and applying machine learning algorithms to it. Data visualization in the explorer panel minimizes visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships. Before the data is loaded and used by the explorer interface, it should be first stored in spread sheet or database and changed in to a name called dataset. Tasks like loading of data, data preprocessing, attribute selection, data visualization and using different learning algorithms, such as classification, clustering and association rule extraction are accessible the interface of the explorer. It allows us to provide a uniform interface to these learning

algorithms, along with methods for pre- and post processing and for evaluating the result of learning schemes on any given dataset.

The Knowledge flow interface is another approach for accessing and using the same functionality with explorer but with a drag-and-drop style of Knowledge flow module. Experimenter interface is used to test and evaluate machine learning algorithms. The last but not the least interface is a simple command line interface which is used as an interface for typing commands (Soman & Bobbie, 2005).

The ways of using Weka by researchers may vary. One way of using Weka is to apply a learning method to a dataset and analyze its output to extract information and knowledge about the data. Another is to apply different learners and compare their performance in order to choose one for prediction (Witten et al., 2000).

Thus, for the purpose of this study, the pre processed data set was loaded on Weka machine learning environment and each of the chosen algorithms were run one by one. For the test options, cross-validation' with number of folds '10' was set. In this approach, the entire data set is divided into 10 mutually exclusive subsets (or folds) or partitions with approximately the same class distribution as the original data set (stratified). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining 9 folds, leading to 10 independent performance estimates. In each of the 10 iterations, 1 fold is used as test (holdout) sample while the remaining 9 are used for model building. For methods comparison studies with relatively smaller data sets, the k-fold types of experimentation methods are recommended. In essence, the main advantage of 10-fold (or any number of folds) cross-validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as a holdout sample (Delen & Sirakaya, 2006). Note down the results/critical outputs on each algorithm were also the main task that was performed in the experimental process. Then, the results of each of the selected algorithms were summarized.

3.3.1.1 Machine understandable format in WEKA

The application of the data to Weka required that some preprocessing be undertaken. The dataset produced in Excel for the statistical processes were copied and then converted to .csv file format to allow them to be applied to Weka. The .csv file extension allowed initial analysis to be conducted, with later conversion to be taken in to an arff format (with .arff extension) data file for the experimental outcome to be saved. In arff data format, the internal name of the data set should be stated using the symbol '@. Attributes should also be defined by preceding the symbol'@' and then with their relevant values and data types. The rest of the dataset consists of the token @data, followed by comma-separated values for the attributes. Fig.3.6 depicts the sample of machine understandable format of the dataset in Weka employed for this study.

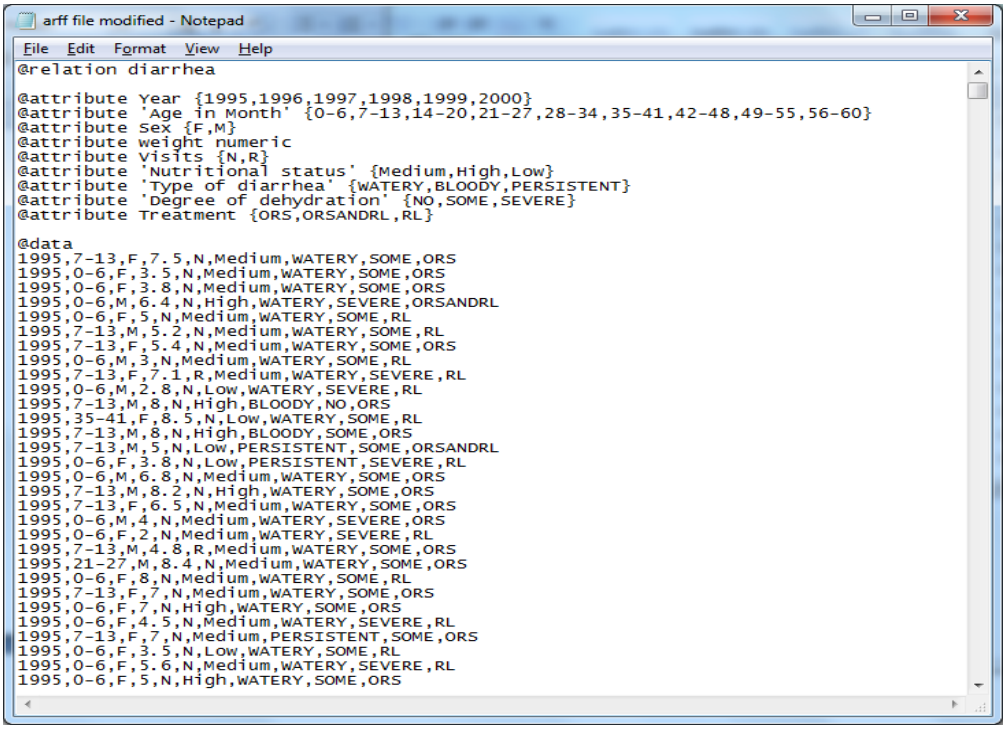

Figure 3.6: Sample machine understandable format of the data set in Weka for the study.

3.3.2 Decision Tree Classifier

Decision trees are a way of representing a series of rules that lead to a class or value (TCC, 2005). They are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. They provide a set of rules that can be applied to a new (unclassified) dataset to predict which records will have a given outcome. They are powerful and popular tools for classification and prediction (Untwal, 2008).

In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. Decision trees are also useful for exploring data and gaining sight into the relationships of a large number of candidate input variable to the target variable. Decision trees can represent rules usually accompanied by the form "if condition then outcome," which constitute the text version of the model that can readily be expressed and humans can easily understand them. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of continuous variable (Hajizadeh et al., 2010).

The decision tree that predicts a categorical outcome, for example, died or alive, is called a "classification tree". If it predicts a continuous variable such as age groups that are affected by a certain disease is called a "regression tree". A decision tree is a hierarchical structure with each node contains decision attribute and node branches corresponding to different attribute values of the decision node. The goal of building decision tree is to partition data with mixing classes down the tree until the leaf nodes contain pure class (Kerdprasop, N. & Kerdprasop, K., 2009).

There are many decision tree algorithms in Weka software. The most common decision tree construction algorithm is J48 which was used for this paper (H. and E. Frank, 2000). J48 is a decision tree learner. It is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, following a path from the root to the leaves of the tree, a sequence of such tests is performed resulting in a decision about the appropriate class of the data set. The decision trees are constructed in a top-down fashion by choosing the most appropriate attribute each time. An information-theoretic measure is used to evaluate features, which provides an indication of the "classification power" of each feature. Once a feature is chosen, the training data are divided into subsets, corresponding to different values of the selected feature, and the process is repeated for each subset, until a large proportion of the instances in each subset belong to a single class (Soman & Bobbie, 2005).Decision trees are a graph of choices or decisions. Each node, from the root node down, represents a decision. The final node of any branch (the leaf node) is used for classification. An example of hypothetical tree is depicted in fig.3.7.

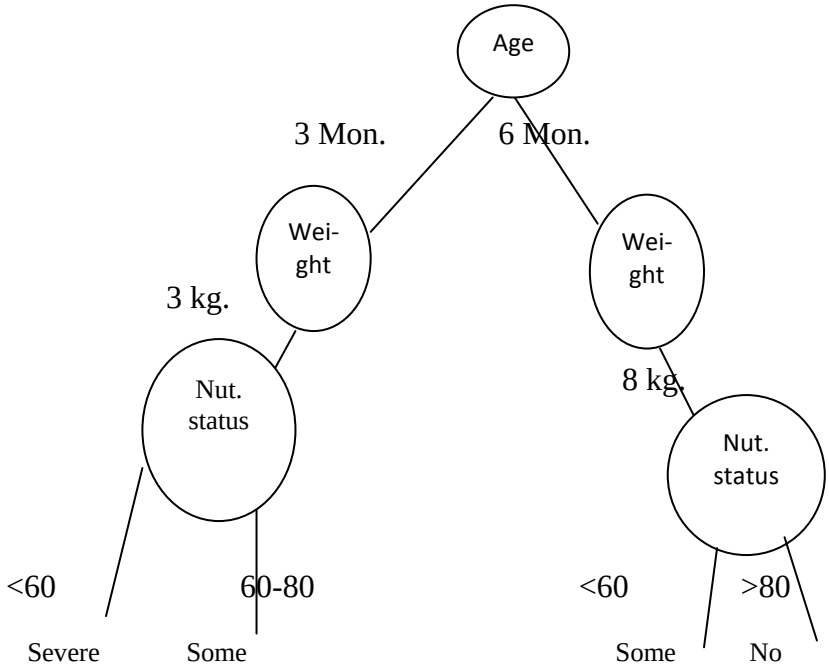

Figure 3.7: An example of hypothetical decision tree with decision points and rules associated with child's age getting diarrhea and degree of dehydration.

(Nut. = Nutritional)

3.3.2.1 Constructing Decision Tree

Decision tree programs construct a decision tree from a set of training sets. The main focus of a decision tree growing algorithm is selecting which attribute to test at each node in the tree. The decision trees are constructed in a top-down fashion by choosing the best and most appropriate attribute each time. An information-theoretic measure is used to evaluate features and select the best attribute, which provides an indication of the "classification power" of each feature. In information -theoretic measure, the concept of Entropy and Information Gain (IG) are used by the algorithm. Information gain (IG) is measured as the amount of the entropy (S) difference when an attribute contributes the

additional information about the class, Where as Entropy(S) is the sum of the probability of each label times the log probability of that same label (El-Telbany et al., 2006).

In order to define information gain precisely, we need to define a measure commonly used in information theory, called entropy, which characterizes the impurity of an arbitrary collection of examples. Given a set S, containing only positive and negative examples of some target concept (a 2 class problem), the entropy of set S relative to this simple, binary classification is defined as:

$$\text{Entropy}(S) = -P_P \log_2 P_P - P_n \log_2 P_n \qquad (1)$$

Where p_p is the proportion of positive examples in S and p_n is the proportion of negative examples in S.

The entropy is 0 if all members of S belong to the same class. For example, if all members are positive (p_p= 1), then p_n is 0, and Entropy(S) = -1´ $\log_2$ (1) - 0´ $\log_2$0 = -1* 0 – 0* $\log_2$0 = 0. The entropy is 1 (at its maximum) when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

The above computation of entropy is in the special case where the target classification is binary. When the target attribute takes values more than two, say k different values, then the entropy of S relative to this k-wise classification is defined as:

$$\text{Entropy}(S) = \sum_{i=1}^{k} -p_i \log_2 p_i \qquad (2)$$

Where p_i is the proportion of S belonging to class *i*. if the target attribute can take on k possible values, the maximum possible entropy is $\log_2$k.

As entropy is a measure of the impurity in a collection of training examples, information gain is a measure of the effectiveness of an attribute in classifying the training data. It is simply the expected reduction in entropy caused by partitioning/splitting the examples according to this attribute. Say, this attribute is considered to be A, the information gain (S, A) of an attribute A, Relative to a collection of examples S, is defined as:

$$\text{Gain}(S, A) = \text{Entropy}(S) \quad - \sum_{V \in \text{Values}(A)} \frac{|S_V|}{|S|} \text{Entropy}(S_V) \qquad (2)$$

where Values(A) is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S \mid A(s) = v\}$). Note the first term in the equation for Gain is just the entropy of the original collection S and the second term is the expected value of the entropy after S is partitioned using attribute A. The expected entropy described by this second term is the sum of the entropies of each subset S_v, weighted by the fraction of examples $|S_v|/|S|$ that belong to S_v. Gain (S,A) is therefore, the expected reduction in entropy caused by knowing the value of attribute A. In another way, Gain(S,A) is the information provided about the target attribute value, given the value of some other attribute A. The value of Gain(S,A) is the number of bits saved when encoding the target value of an arbitrary member of S, by knowing the value of attribute A.

According to Hamilton et al. (2011), the process of selecting a new attribute and partitioning the training examples is repeated for each non-terminal descendant node, this time using only the training examples associated with that node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree.

3.3.2.2 Rule induction

Rule induction is the process of extracting useful ‘if then’ rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. It has the form:

IF conditions THEN conclusion

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction (Kaur & Wasan, 2006).

3.3.3 Naïve Bayes (NB) Classifier

Naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions (Bhargavi & Jyothi,2009).That is, there are no dependence relationship among the attributes given the value of the class variable (Singh, 2009). Despite this strong assumption, the algorithm tends to perform well in many class prediction scenarios. Experimental studies suggest that Naïve Bayes tends to learn more rapidly than most induction algorithms. Given a new instance, the classifier estimates the probability that the instance belongs to a specific class, based on the product of the individual conditional probabilities for the feature values in the instance. The exact calculation uses Bayes theorem and this is the reason why the algorithm is called a Bayes classifier (Soman & Bobbie, 2005).

In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a disease might be considered to be severe if the patient depicts dramatic

loss of body weight, high fever, decrease in food intake and others. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this disease is severe. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In spite of their Naive design, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. The Naive Bayesian classifier is fast and incremental, and can deal with discrete and continuous attributes with excellent performance in real-life problems. It has capability to solve also non-linear problems while retaining all advantages of Naive Bayes (Bhargavi & Jyothi, 2009).

Learning a Naive Bayes classifier is straightforward and involves estimating the probability of attribute values within each class from the training instances. Probabilities are estimated by counting the frequency of each discrete attribute values. For numeric attributes, it is common practice to use the normal distribution (Singh, 2009). Given a new instance, the classifier estimates the probability that the instance belongs to a specific class, based on the product of the individual conditional probabilities for the feature values in the instance. The exact calculation uses Bayes theorem and this is the reason why the algorithm is called a Bayes classifier.

The main advantage of using Naïve Bayes is that they are probabilistic models, robust to noise found in real data. The Naive Bayes classifier presupposes independence of the attributes used in classification. However, it was tested on several artificial and real data sets, showing good performances even when strong attribute dependences are present. In addition, the Naive Bayes classifier can outperform other powerful classifiers when the sample size is small (Ferrari, 2005).

3.3.3.1. Naïve Bayesian Classification Algorithm

The Naïve bayes classifier can predict class membership probabilities, such as the probability that a given sample belongs to a particular class (Subbalakshmi et al., 2011). This is performed by Naïve Bayes classifier as follows.

Let T be a training set of samples, each with their class labels.There are k classes, C_1,C_2, . . . ,Ck. Each sample is represented by an n-dimensional vector, $X = \{x_1, x_2, \ldots, x_n\}$, depicting n measured values of the n attributes, A_1,A_2, . . . ,A_n, respectively.

Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class C_i if and only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$.

Thus we find the class that maximizes $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (1)$$

As P(X) is the same/constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class priori probabilities, $P(C_i)$, are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots = P(C_k)$, and we would therefore maximize $P(X|C_i)$. Otherwise we maximize $P(X|C_i)P(C_i)$. The class prior probabilities may be estimated by $P(C_i)=|C_i,T|/|T|$**,** where $|C_i,T|$ is the number of training samples of class C_i in T.

Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$.

In order to reduce computation in evaluating $P(X|C_i)$, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample (i.e., there are no dependence relationships among the attributes). Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \qquad (2)$$

$$= P(x_1|C_i) \times P(x_2|C_i) \times \ldots P(x_k|C_i)$$

The probabilities $P(x_1|C_i)$, $P(x_2|C_i)$, . . . , $P(x_k|C_i)$ can easily be estimated from the training set. Recall that here x_k refers to the value of attribute A_k for sample X.

(a) If A_k is categorical, then $P(x_k|C_i)$ is the number of samples of class C_i in T having the value x_k for attribute A_k, divided by (C_i, T), the number of sample of class C_i in T.
(b) If A_k is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean μ and standard deviation σ defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad (3)$$

So that

$$P(x_k|C_i) = g(x_k, \mu C_i, \sigma C_i)$$

We need to compute μC_i and σC_i , which are the mean and standard deviation of values of attribute A_k for training samples of class C_i.

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

CHAPTER FOUR

EXPERIMENTATION

4.1 Experimental Setup Overview

The data analysis and classification was carried out using Weka software environment. The data set collected from the diarrheal disease control and training center in Tikur Anbessa Hospital consisted of 5,572 patients' records and initially intended for this study. However, after 'SMOTE' was run to balance the size of class labels, the number of instances were over sampled and reached to 13,710 and 16,460 for the target attribute classes of 'Treatment' and 'Type of diarrhea' respectively. Therefore, the size of records obtained after the class labels were balanced was used for the study. Nine features were identified by the researchers which were deemed to be pertinent to the study. The attributes identified are year, age, sex, weight, visits, nutritional status, type of diarrhea, degree of dehydration and treatment.

As discussed in chapter 3, section 3.3, the package employed for the purpose of this research is the Weka software and J48 decision tree and Naïve Bayes classifiers. Weka provides three options to partition the dataset in to training and test data. These are: preparing distinct files for training dataset and test dataset; cross validation with possibility of setting variety number of folds (the default was 10 fold) and percentage split. A 10-fold cross validation has been used for this research. This test option was selected for this study with the intention to be free from bias during dataset partitioning for training and testing. As it is explained in chapter 3, section 3.3.1, in cross validation mode, the data is divided into some number of partitions of the data, in this case, 10 approximately equal proportions, and each in turn was used for testing while the remainder was used for training. This process repeats 10 times and at the end, every instance has been used exactly once for testing. Finally the average result of the 10 fold

cross validation is considered (Witten et al.,2000).Therefore, from a total records of 13,710, which were used for model building using 'Treatment' as a target class , 12,339 instances (90%) of these records were used to build(train) models and the remaining 1371 (10%) of the dataset were used to test the performance of the models. On the other hand, from a total of 16,460 records which were used for model building using 'Type of diarrhea' as a target class,14,814(90) of these records were used to build(train) models and the remaining 1,646 (10%) of the dataset were used to test the performance of the model. The models were also evaluated and compared for their classification and prediction performance.

4.2 Defining the Target Attributes

The classifiers in Weka are designed to be trained to predict a single 'class' attribute, which is the target for prediction. Some classifiers can only learn nominal classes; others can only learn numeric classes; still others can learn both. Weka classifiers need predefined classes in order to train and build classification models (Bouckaert et al., 2010).Unless they are given the target attribute by the data miner, the last attribute is taken as a target class by default. It is possible to choose any attribute name as a target class no matter its position in the list while running the program. Therefore, the training attribute should be pre-classified so that the data mining algorithms know what the user is looking for. Therefore, 'Treatment' and 'Type of diarrhea' were used as a target class for this study. The selection of the target attribute was made using the input from domain experts. These class attributes were selected due to their appropriateness in indicating prevalence of a specific clinical type of diarrhea.

4.3 Model Building and Result Analysis

4.3.1 Decision Tree Model Building

A decision tree is a classifier in which previously unobserved records can be fed into the tree. At each node it will be sent either left or right according to some test. Ultimately, it will reach a leaf node and be given the label associated with that leaf. From the results of the decision tree classifier, it is possible to generate interesting rules. In fact, decision tree methods are often chosen for their ability to generate understandable rules in addition to their classification and prediction capabilities. Using J48 Decision Tree classifier, two models were performed for each target class. The first one was with the default settings of the program and the other by modifying 'minNumObj' and 'confidenceFactor' parameters. The process of classifying records proceeds until the number of records at each leaf reached to the value given using 'minNumObj' where as the 'confidence factor' parameter is used for pruning. These figures were set after a number of trials were made.

The first experiment provided with better accuracy as compared to the second. However, the decision tree generated from the first experiment was very large, complex and difficult to generate understandable rules. The size of a tree and number of leaves produced from this training was 636 and 393 respectively. But, if one needs to pass through all the nodes of this tree in order to come out with valid rule sets, it is cumbersome and difficult. Thus, the researcher modified values of the aforementioned parameter settings of the program in order to minimize the size of the tree and number of leaves. For this purpose, the parameters 'minNumObj (minimum number of instances per leaf)' and 'confidence factor' were set to 25 and 0.025 which were 2 and 0.25 by default respectively.

The modification of these parameters has reduced the size of the tree and number of leaves from 636 and 393 to 212 and 139 respectively. The size of the tree and number of leaves of the second trial was also large. In fact, it was possible to reduce the size of the

tree and number of leaves below this size. However, further modification of the afore mentioned parameter settings to reduce the size of the tree and number of leaves below the size mentioned above resulted in decreasing overall accuracy of the model.

Similarly, two experiments were executed using 'Type of diarrhea' as a target class. The first model which was performed with the default setting of the program provided better accuracy as compared with the second that was carried out by modifying the same parameter settings mentioned above. However, the size of a tree and number of leaves generated from the first experiment were also very large and complex. The size of the tree and number of leaves were 1163 and 747 respectively. Therefore, it was necessary to decrease the size of the tree and number of leaves so as to come up with more understandable decision tree structures. To this end, 'minNumObj' and 'confidence factor' parameters were set with the same values as above. After these parameter settings were modified, the size of the tree and number of leaves reduced to 366 and 250 respectively. All trials performed to minimize the size of the tree and number of leaves below this resulted with very low accuracy rate.

Hence, the researcher selected the results of the second trials for both of the target classes as a working model for further stages of the study. The selected rules and results of a confusion matrix from the selected trials are presented below. Results which show overall accuracy, size of trees and number of leaves for some trials carried out using Decision tree classifier before and after modification of the aforementioned parameters are presented in appendix part.

4.3.1.1 Generating Rules from Decision Trees

One of the advantages of classification using decision tree classifiers is to generate set of rules. Decision rule algorithms classify records by following a set of classification directives

or rules. The rules indicate whether a given class is a good indicator or not that the feature belongs to a given category. The rules may be combined in the form of a complex decision tree. A rule is a correlation found between the main variable (dependent) and the others (independent variables).This produces rules that are clear in that it doesn't matter in what order they are executed (Shailja,2009).

Below, Fig.4.1 and 4.2 present the selected rules extracted from models which were built after parameter settings modification.

Selected Rules using 'Treatment' as a target class

1. IF DoD=No,
   Then Treatment: ORS (832.0/47.0)
2. IF DoD=Some and NS=Medium and Age=7-13 and ToD=Watery and Year of Treatment = 1998,
   Then Treatment: ORS (164.0/32.0)
3. IF DoD=Some and NS=Low and Year of Treatment=1995 and Age=7-13,
   Then Treatment: ORS and RL (72.0/17.0)
4. IF DoD=Severe and Year of Treatment =1995 and ToD= Persistent and Age=7-13,
   Then Treatment: RL (233.0/7.0)
5. IF DoD=Severe and Year of Treatment=1995 and ToD=Bloody,
   Then Treatment: ORS and RL (99.0/17.0)
6. IF DoD=Severe and Year of Treatment=2000,
   Then Treatment: RL (529.0/50.0)
7. IF DoD= SEVERE and Year of Treatment = 1995 and ToD= Watery and Age=14-20 and Weight <= 10.170634,
   Then Treatment: RL (60.0/3.0)

8. IF DoD= Severe, Year of Treatment = 1995 and ToD= Watery and Age=14-20, Weight > 10.170634,

   Then Treatment: ORSANDRL (72.0/10.0

9. IF DoD=Some and NS= Low, Year of Treatment = 1995 and Age=0-6 and Weight >3.724652,

   Then Treatment: ORS and RL (114.0/30.0)

10. IF DoD=Some,NS= Low, Year of Treatment = 1995and Age=7-13, Weight > 6.449342,

    Then Treatment: ORS and RL (72.0/17.0)

Figure 4.1: Selected rules extracted from J48 decision tree classifier based on 'Treatment' as a target class.

(Note: DoD =Degree of Dehydration, ToD =Type of Diarrhea, NS= Nutritional Status)

The selected rules presented in fig. 4.1 indicate the possible conditions in which a treatment for diarrhea could be classified in each class. Degree of dehydration, type of diarrhea, year, nutritional status, age and weight are the features involved to determine the type of treatment given for patients with diarrhea. Although the classifier included all the features of the dataset in generating decision trees, some might not be involved for all individual rules. As it is indicated also from these rules, degree of dehydration is the basis for classification of treatment modalities. These rules indicate useful correlations among the features in classifying class labels. For instance, one can see from rule1 that if degree of dehydration is detected to be 'NO' among patients with diarrhea, ORS (Oral Rehydration Salt) is a treatment that should be given. As rule 6 indicates, RL could be given for patients with any type of diarrhea, if they are severely dehydrated. Rule 3 also indicates that diarrheal patients with age of 7-13 months should be treated with both ORS and RL, if their degree of dehydration and nutritional status are detected as 'some' and 'low' respectively. From rule 4 and 5 one can also observe that RL is given for diarrheal patients who are severely dehydrated due to persistent and bloody diarrhea.

Table 4.1: Output from J48 Decision Tree classifier based on 'Treatment' as a target class.

		Predicted				
		ORS	ORS AND RL	RL	**Total**	**Score (Accuracy rate)**
Actual	ORS	3826	205	303	**4334**	**88.27%**
	ORS AND RL	136	4760	160	**5056**	**94.14%**
	RL	431	368	3521	**4320**	**81.50%**
	Total	**4393**	**5333**	**3984**	**13,710**	**88.30%**

The above values of the confusion matrix in table 4.1 depicts that out of the total 13,710 data provided to the program, 12,107(88%) records were classified correctly and the remaining 1,603(12%) were classified incorrectly. The results of this table also indicates that 205 and 303 records from actual class ORS were classified as RL and, ORS and RL classes respectively, while 136 and 431 records from class ORS and RL, and RL are wrongly classified as ORS class. On the other hand, 368 records were misclassified as ORS and RL class, while they are actually be RL class. These results also depict that 160 records were wrongly classified as RL class while they are actually be ORS and RL class.

One can also see from table 4.1 that most of the patients were treated with both ORS and RL as compared to ORS or RL alone. That is, the model predicted that many of the patients who come to the center provided with both ORS and RL, and regardless of the order they were given. According to the suggestions of the domain expert for these results, most diarrheal patients initially might not be severely dehydrated and malnourished. Thus, if the degree of dehydration was found to be 'no' or 'some' ORS is a common intervention for treating those patients. But, if the patients could not be

recovered soon, and their degree of dehydration as well as nutritional status could not be improved, RL was used as an intervention. The domain expert also said that patients might be severely dehydrated and malnourished when they come to the center. For such occasions, RL could be given directly. When patients indicated improvements on their health conditions, ORS could be given in order to replace lost fluid from their body. Therefore, there might be a probability for most patients to be treated with both ORS and RL.

Selected Rules using 'Type of diarrhea' as a target class

1. IF Treatment=ORS and Year of Treatment=1997 and DoD= Some,
   Then ToD: Watery (612.0/13.0)
2. IF Treatment = RL and Year of Treatment =1997 and Age=0-6 and DoD=No,
   Then ToD: Watery (1.0)
3. IF Treatment = ORS and RL and Age=7-13 and DoD= Severe and Weight <= 6.999723,
   ThenToD: Persistent (169.0/7.0)
4. IF Treatment =ORS and Year of Treatment =1996 and DoD= Some and Age=14-20 and NS= Low, Weight <= 7.999821,
   Then ToD: Persistent (187.0/3.0)
5. IF Treatment = ORS and RL and Age=0-6 and Weight<= 4.409932,
   Then ToD: Persistent (182.0/24.0)
6. IF Treatment = ORS and RL and Age=7-13 and DoD= Severe and Weight <= 6.999723,
   Then ToD: Persistent (169.0/7.0)
7. IF Treatment = RL and Year of Treatment =1995,
   Then ToD: Persistent (764.0/92.0)
8. IF Treatment = RL and Year of Treatment =1996 and NS= Low,
   Then ToD: Persistent (248.0/18.0)

9. IF Treatment = RL and Year of Treatment =1997 and Age=7-13 and NS= Medium and Weight < 7.72601,
   Then ToD: Persistent (386.0/23.0)
10. IF Treatment = RL and Year of Treatment =1997 and Age=0-6 and DoD=Severe
   Then ToD: Bloody (551.0/65.0)
11. IF Treatment =RL and Year of Treatment =2000 and Age=0-6 and Weight <= 4.999472,
   Then ToD: Bloody (184.0/26.0)
12. IF Treatment =RL and Year of Treatment =1998 and NS= Low,
   Then ToD: Bloody (442.0/79.0)
13. IF Treatment =RL and Year of Treatment =2000 and Age=7-13 and Weight < 7.329824,
   Then ToD: Bloody (117.0/15.0)
14. IF Treatment =RL and Year of Treatment =2000 and Age=28-34,
   Then ToD: Bloody (6.0/2.0)
15. IF Treatment =RL and Year of Treatment =2000 and Age=42-48,
   Then ToD: Watery (1.0)
16. IF Treatment =ORS and Year of Treatment =1997 and DoD=No,
   Then ToD: Watery (128.0/1.0)

Figure 4.2: Selected rules from J48 decision tree classifier based on 'Type of diarrhea' as a target class.

The above rules indicated in fig.4.2 reveal how a target class 'Type of Diarrhea' is classified based on selected attributes. The selection was made by the classifier itself. It has included 8 attributes to build a model. The only variable excluded by the classifier was 'Visits'. Hence, instances were classified into the predefined classes of diarrheal type using these attributes. In fact, in classifying diarrheal disease, 'treatment' has been the

bases for this classification. Year, age, sex, nutritional status, degree of dehydration and weight are also other features involved in building the decision tree by the classifier.

Important and interesting findings can be identified from the above rules. For example, rule 6 show that the type of diarrhea is considered to be persistent among patients of age 7-13 months, if their weight is less than 7 kg, degree of dehydration is severe and treated with both ORS and RL. Rule 10 also depicts that a type of diarrhea is found to be bloody, if dehydration level is severe among patients of age 0-6 months and RL is taken as an intervention. As it is also shown in rule 16, the type of diarrhea was considered to be watery, if patients are not dehydrated at all (No- dehydration) and ORS is given as a treatment. Rules 11, 13, 14 and 15 revealed that the possibility to be attacked by diarrhea is reducing as age increases. Thus, the rules can be used to justify the features and values which classified diarrheal type in to one of the predefined classes.

In general, from these rules, even though there may be some less important relationships among the features involved in the model, most of them are interesting and useful. Therefore, health care workers and other concerned bodies can benefit from these rules to facilitate and improve the services being given and to make good decisions concerning diarrheal disease.

Table 4. 2: Output from J48 Decision Tree classifier based on 'Type of diarrhea' as a target class.

		Predicted				
		Watery	Bloody	Persistent	**Total**	**Score (Accuracy Rate)**
Actual	Watery	3616	867	553	**5036**	**71.80%**
	Bloody	468	5044	216	**5728**	**88.05%**
	Persistent	130	129	5437	**5696**	**95.45 %**
	Total	**4214**	**6040**	**6206**	**16,460**	**85.64%**

The above values of the confusion matrix in table 4.2 revealed that out of the total 16,460 records used to build a model, 14,097(86%) were classified correctly and the remaining 2363(14%) classified incorrectly. One can also see that 867 and 553 records were incorrectly classified as bloody and persistent classes while they should be watery class. On the other hand, 468 and 130 instances were classified wrongly as watery class while they actually be bloody and persistent classes respectively. Results from this table also show that 129 and 216 records were incorrectly classified as bloody and persistent classes while the actually be persistent and bloody respectively.

There might be various reasons for the misclassification from one type of diarrhea to another type of diarrhea. One possible reason is that the model was selected on the basis of the soundness of the decision trees and rules that it generates rather than its classification performance. The researcher also cross checked the relations among some misclassified sample records from the output predictions to find possible reasons. One possible reason found is that many patients who severely dehydrated were treated with ORS. Moreover, although the nutritional status of most patients was found to be medium or low, ORS alone was used as an intervention.

One can also see from the results in table 4.2 that the model performed quite well for persistent class. That is, prevalence of persistent diarrhea was best predicted by the model. However, the performance of the model to predict watery diarrhea was poor. Bloody diarrhea was well predicted as compared with watery diarrhea.

4.3.2 Naive Bayes (NB) Model Building

The second data mining technique employed in this study was Naive Bayes classifier. To build this model, the same software package (Weka software) that was used for decision tree model building is employed. The test option used in this experiment was also 10-fold cross validation. As it is discussed in chapter 3 section 3.4, so as to start building a model to a specific dataset, there is usually a need to prepare the dataset in a form which is suitable for the particular data mining technique and software. An attempt has been made to clean and preprocess the data for J48 decision tree classifier. All of the tasks carried out before are also applicable for NB classifier.

In order to carry out an experiment (classification of records of this dataset) using the Naive Bayes classifier of the Weka software, the researcher used the same dataset and target classes which are employed for building decision tree model so far. Experiments with this classifier were also conducted with and without modification of parameter settings for both target classes. That is, first, with default parameters of the program and then by modification of 'usesupervised descritization' parameter from its default value of 'False' to 'True' value. This parameter is used to convert numeric attribute values to nominal ones. Because, there is one attribute (weight) with numeric data type from the data set used in this study. This was aimed for changing this attribute values to nominal and inspecting whether there was improvement or not on the output generated from this classifier. According to Ceci (2005), the Naive Bayes classifier can be more successful for discrete attributes.

The modification of the aforementioned parameter settings improved the accuracy performance of Naïve Bayes classifier for both target classes as compared with the accuracy resulted from the default settings of the program. The model provided with an accuracy of 68% and 80% before and after the 'usesupervised descritization' parameter was modified respectively, when 'Treatment' was used as a target class. On the other hand, when 'Type of diarrhea' was used as a target class, the accuracy was improved from 59% to74% after the same parameter was modified. Therefore, the results of the model with improved accuracy were selected for the purpose of this study. The confusion matrix output of the selected models for 'Treatment' and 'Type of diarrhea' classes are depicted below in table 4.3 and 4.4.

Table 4. 3: Output of Naive Bayes classifier based on 'Treatment' as a target class.

		Predicted				
		ORS	ORS AND RL	RL	**Total**	**Score (Accuracy rate)**
Actual	ORS	3885	164	285	**4334**	**89.64%**
	ORSAND RL	445	4037	574	**5056**	**79.84%**
	RL	313	1023	2984	**4320**	**69.07%**
	Total	**4643**	**5224**	**3843**	**13,710**	**79.54%**

As it is shown in table 4.3, out of the total 13,710 records used for building a model using NB classifier, 10,906(80%) of records were classified correctly where as 2804(20%) were classified incorrectly. Results in table 4.3 also depict that 164 and 285 of records were misclassified as ORS and RL, and RL classes respectively while they actually be ORS class. On the other hand, 445 and 313 instances were misclassified as ORS class while they should be ORS and RL, and RL classes respectively. 574 instances

were misclassified as RL class while they actually be ORS and RL class.1023 records were also misclassified as ORS and RL class while they actually be RL class.

As one can see the results of this training scheme, ORS was best classified as a treatment for patients with diarrhea as compared with RL and both. According to this model, using RL to treat patients with diarrhea was classified with an accuracy rate below other treatment classes.

Table 4. 4: Output of Naive Bayes Classifier based on 'Type of diarrhea' as a target class.

		Predicted				
		Watery	Bloody	Persistent	**Total**	**Score (Accuracy**
Actual	Watery	4152	556	328	**5036**	**82.44%**
	Bloody	891	3678	1159	**5728**	**64.21%**
	Persistent	454	900	4342	**5696**	**76.22%**
	Total	**5497**	**5134**	**5829**	**16,460**	**73.94%**

The results depicted in table 4.4 show that 12,172(82%) records from a total of 16,460 records were classified correctly. That means, 4288(18%) of records are incorrectly classified. The classification performance of this classifier was better for class 'watery'. But it was not good in classifying other classes such as 'bloody' and 'persistent'. Especially, in classifying of class 'bloody' it was quite poor. The confusion matrix result in table 4.4 also indicates that 556 and 328 records were misclassified as bloody and persistent respectively ,while they are actually be watery. On the other hand, 891 and 454 of records were misclassified as watery while they are actually be bloody and persistent

respectively. Table 4.4 also indicates that watery diarrhea was well predicted by Naïve Bayes classifier.

4.4 Performance Evaluation of J48 Decision Tree and Naive Bayes Classifiers

One of the objectives of this study was to compare and evaluate the techniques which were used in the study, such as decision tree and Naïve Bayes classifiers and to select the one, which performs the best. To evaluate and compare the performance of each of the classifiers involved in this study, the standard metrics of accuracy, precision, recall, F-measure, True-positive and False-Positive Rates were used. Time taken by each classifier to build the selected models, and number of instances which classified correctly and incorrectly were also other parameters used to compare and evaluate classifiers' performance. The true positive and false positive rates, precision, recall and f-measure, values of each classifier for each class label were used for evaluation.

Most of these metrics were calculated using the predictive classification table (Confusion Matrix). Confusion matrix columns indicate the predicted class (classified), and the rows, the existing classes (real). They show how many instances of each class have been assigned to each class. Diagonal matrix indicates correct predictions and other elements of the matrix show incorrect predictions (FP+FN). The True Positives (TP) measures how many instances of a given class are correctly classified. The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. In the confusion matrix, this is the diagonal element divided by the sum over the relevant row. It is equivalent to Recall. The False Positives (FP) measures how many instances of other classes are confused with a given class. It is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. In the matrix, this is the

column sum of class x minus the diagonal element, divided by the rows sums of all other classes.

The number of correctly predicted values in the total number of predicted values is stated in the precision parameter that indicates that the proposed model has predictive power and is conclusive. It is the proportion of the examples which truly have class x among all those which were classified as class x. In the matrix, this is the diagonal element divided by the sum over the relevant column. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified instances (Xhemalid et al**,** 2009). F-measure is a way of combining recall and precision scores into a single measure of performance (Kumar & Rathee, 2011).

The evaluation was performed on the results of the aforementioned parameters for the selected model of both the target classes used in the experiment. The results of these models indicated that the classification task of records using 'Treatment' as a target class for both decision tree and Naïve Bayes have performed well as compared with using 'Type of diarrhea' as a target class. Decision tree has shown an accuracy rate of about 80%, while Naïve Bayes classified about 80% of records correctly. In the classification task of records using the target class of 'Type of diarrhea', the accuracy of Decision Tree was about 86% where as that of Naïve Bayes was 74%.Most of the results of the parameters, such as precision, recall and f-measure also indicated that Decision Tree outperformed than Naïve Bayes classifier. All these results revealed that the decision tree performed better. Hence, it is reasonable to conclude that the decision tree data mining technique is more appropriate and preferable to this particular case than the Naïve Bayes. The results of the aforementioned parameters which were used to compare classifiers' performance using 'Treatment' and 'Type of diarrhea' as a target class are depicted below in table 4.5 and 4.6.

Table 4. 5: Predictive Performance of Classifiers based on 'Treatment' as a target class.

Parameters		Classifiers	
		J48 Decision Tree	Naïve Bayes
Over all Accuracy		88.3%	79.54%
Precision	ORS	87.1%	83.7%
	ORS AND RL	89.3%	77.3%
	RL	88.4%	77.6%
Recall	ORS	88.3%	89.6%
	ORS AND RL	94.1%	79.8%
	RL	81.5%	69.1%
F-measure	ORS	87.7%	86.6%
	ORS AND RL	91.6%	78.5%
	RL	84.8%	73.1%
True-positive Rate	ORS	88.3%	89.3%
	ORS AND RL	94.1%	79.8%
	RL	81.5%	69.1%
False-positive Rate	ORS	6%	8.1%
	ORS AND RL	6.6%	13.7%
	RL	4.9%	9.1%
Correctly classified instances		12,107	10,906
Incorrectly classified instances		1603	2804
Time to build model (in sec.)		0.23	0.16

Table 4.5 shows the Accuracy, Precision, Recall and F-Measure results achieved by the Naïve Bayes and Decision Tree classifiers. The table also depicts the True-positive and False-positive rates, time taken in building models and number of correctly and incorrectly classified instances by each classifier. These results show that Decision Tree classifiers outperformed the Naïve Bayes classifier for most of the metrics used for

evaluation. However, the time taken to build a model by the Naïve Bayes classifier was smaller than Decision tree classifier. That is, the Naïve Bayes classifier tends to learn more rapidly for the given dataset.

As one can see the accuracy rate of both models, Decision Tree achieved 88% while Naïve Bayes scored 80%.This indicates that the number of diarrheal patients who really treated and not treated by the specified treatment modalities were best classified and predicted by the decision tree classifier. The precision, recall and f-measure results show that except for recall or True-positive rate of class lable 'ORS', the Decision Tree classifier outperformed Naïve Bayes classifier for all other results of parameters.

Table 4. 6: Predictive Performance of Classifiers based on 'Type of diarrhea' as a target class.

Parameters		Classifiers	
		J48 Decision Tree	Naïve Bayes
Over all Accuracy		85.64%	73.94%
Precision	Watery	85.8%	75.5%
	Bloody	83.5%	71.6%
	Persistent	87.6%	74.5%
Recall	Watery	71.8%	82.4%
	Bloody	88.1%	64.2%
	Persistent	95.5%	76.2%
F-measure	Watery	78.2%	78.8%
	Bloody	85.7%	67.7%
	Persistent	91.4%	75.3%
True Positive-Rate	Watery	71.8%	82.4%
	Bloody	88.1%	64.2%
	Persistent	95.5%	76.2%
False positive-Rate	Watery	5.2%	11.8%
	Bloody	9.3%	13.6%
	Persistent	7.1%	13.8%
Correctly classified instances		14,097	12,172
Incorrectly classified instances		2363	4288
Time to build model (in sec.)		0.42	0.22

As results from table 4.6 depict, the Decision Tree classifier classifies the dataset with an accuracy of 86% while Naïve Bayes classifies with an accuracy of 74% .As one can see also results of precision, recall and f-measure of each class label, Decision Tree outperformed the Naïve Bayes classifier for most of these parameters. Similarly, results

of True-positive and, False-positive rates, number of correctly and incorrectly classified instances show that Decision tree was better classifier as compared with Naïve Bayes. However, in case of time taken to build a model and classification of the individual class ‘watery’, Naïve classifier was better than Decision Tree classifier. Overall, we can say from these results that even if the Naïve Bayes classifier tends to learn more rapidly for the given dataset, J48 Decision Tree classifier outperformed Naïve Bayes for most of the parameters which were used for performance evaluation of these classifiers. Therefore, regardless of training speed, the Decision Tree classifier is more appropriate than Naïve Bayes classifier for this particular data.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The purpose of this study was to investigate the potential applicability of data mining techniques in exploring the prevalence of diarrheal disease. In doing so, 5,572 sample data were initially collected from the diarrheal disease control and training in Tikur Anbessa Hospital. However, the size of class labels for both selected target classes was not balanced for model building process. Therefore, 'SMOTE' was used from the preprocess package of weka to balance those class labels. This increased the size of the sample data from 5,572 to 13,710 and 16,460 for 'Treatment' and 'Type of diarrhea' target classes respectively. The particular data for the research was taken from the data stored in 1995 to 2000 E.C. The total numbers of attributes used in the study were 9. The study was conducted based on the data mining steps or processes discussed in chapter 3. Data collection was a major task which took more of the time of the research. This is due to the manual/paper based format of the original data. Because, it was the first task for the researcher to transform the paper based data to electronic/computer based format.

The methodology employed consisted of steps such as identifying data sources and business understanding, data understanding, data preparation, model building and testing. However, since a data mining task is a cyclic process, these steps were not followed strictly in forward order only. Rather, there was a need to go back and forth among these different steps.

To build models with both classifiers, two attributes such as 'Treatment' and 'Type of diarrhea' were used as target classes. The models were also evaluated and compared for their classification and prediction performance as well as the soundness of the rules extracted from decision trees generated and the best performing models of these

classifiers were then chosen. Both the classifiers were built by using the attributes such as 'Year', 'Age', 'Sex', 'Weight', 'Visits', 'Nutritional Status', 'Type of Diarrhea', 'Degree of Dehydration' and 'Treatment'. A 10-fold cross validation was used as a test option in the model building process.

Various experiments were made iteratively by making adjustments on parameter settings of a program to come up with more understandable and meaningful results and models. When 'treatment' was used as a target class for the Decision Tree approach, the model with more understandable decision trees and rules was identified for the training made by modification of the default parameters. However, there was some reduction in accuracy performance as compared with the model performed with default settings of the program. The parameters 'MinNumobj' and'confidence factor' were modified from their default values of 2 to 25 and 0.25 to 0.025 respectively. At the beginning of the model using the default parameters of the program, this model had an accuracy rate of 91% which is more than from an accuracy of 88% that resulted due to parameter settings modification. However, the structure of the decision tree was very large and difficult to understand. By modifying the aforementioned parameters, it was possible to reduce the size of the tree and number of leaves from 393 and 636 to 139 and 212 respectively. Therefore, the second trial was used as the working model for this study.

When an attribute 'Type of diarrhea ' was used as a target class in building the model by default settings of the program, the structure of the tree was also very large ,complex and difficult to understand. Therefore, the parameters mentioned above were modified so as to reduce the size of the tree and number of leaves for this model. After modification, the size of the tree and number of leaves were minimized from 1163 and 747 to 366 to 250 respectively. However, the accuracy was reduced from 91% to 86% when parameter settings were modified. In spite of its reduction in accuracy, the second trial was selected for further stages of the study due to its simple and understandable decision tree structures as compared with the second trial.

For the Naïve Bayes approach, the classification accuracy performance also has shown some variations before and after parameter settings modification. The accuracy performance of this classifier was improved from 68% to 80% after 'usesupervised descritization' parameter was modified, when 'Treatment' was used as a target class. Similarly, the accuracy was improved from 59% to 74% when 'Type of diarrhea' was used as a target class.

The prediction performance of Decision Tree and Naïve Bayes classifiers also varied for both target classes. The results of this study demonstrate that Decision trees have shown the best accuracy performance in both selected models for 'Treatment' and 'Type of diarrhea' target classes. However, different class labels were best predicted by each classifier. ORS and RL class as a treatment was best predicted by the Decision Tree classifier with an accuracy rate of 94%, where as ORS was best predicted by Naïve Bayes with an accuracy rate of 90%. Among the clinical types of diarrhea, Persistent diarrhea was best predicted by the Decision Tree classifier with an accuracy rate of 95%, while watery diarrhea was best predicted by Naïve Bayes classifier with an accuracy rate of 82%.This indicates that decision trees were the best techniques for the classification of these targeted attributes.

In addition, the comparison of the results of the decision tree and Naïve Bayes models showed that decision tree outperformed Naïve Bayes classifier by the results of most evaluation parameters. This indicated that Naïve Bayes could not perform well for a data set with variables which are dependent each other. Therefore, the decision tree classifier is more appropriate for classification and prediction purposes of this particular clinical data used in the study.

In general, the results from this study were helpful and encouraging. The encouraging results obtained from this study indicate that data mining is really a technology that should be considered to support the health care sector. Data mining can improve not only

to increase accurate diagnosis and successful disease treatment, but also to enhance safety by reducing treatment-related errors as well as to predict future patient behavior and to improve treatment programs. Moreover, understanding the specific prevalent types of a disease can assist concerned bodies in making decisions to abort and control its prevalence before onset and affecting the susceptible groups. By identifying high-risk patients, clinicians can better manage the care of patients today, so they do not become the problems of tomorrow.

5.2 Recommendation

Although this research work is conducted mainly for academic purpose, the researcher believes that the findings of this study can be used for further exploration and investigation of clinical data of diarrheal disease by concerned bodies and organizations. That means, application of data mining technology in clinical data of diarrhea disease is an important research area so as to improve the services being provided as well as to come up with solutions for this disease prevalence. Moreover, the research work can contribute a lot towards a comprehensive study in all other health care sectors in the future.

In the way of doing this study and on the basis of the findings of the research work, the researcher has come up with a sort of tasks that need more attention for the future work.

Thus, the researcher makes the following recommendations based on the results of this study.

- At the beginning of this study, the data was transformed from the paper based file to electronic format, and it was the tedious task that consumed most of the time of the study. Hence, health care institutions including the center from where this data was collected should try to store their day to day data in a computer system. It is important not only for the researchers but also for the health care workers and other concerned bodies when they need to retrieve data. Moreover, inconsistency of values and other features in data collection process were the main constraints in undertaking of the study. Therefore, data storing process needs a regular follow up by the concerned bodies of each health care institution.

- Since this study has used a small percentage of the data which comprises only a six years data of diarrhea in the center to build Naïve Bayes and decision tree

models, it is better to build more comprehensive models by using more additional data from various sources.

- Although encouraging results were obtained from this study, particularly, using Decision Tree classifier, there might be a probability to obtain more accurate and better performing results using other classification and prediction techniques which were not used by the researcher due to time constraint. Therefore, it is recommended that these classifiers should be applied and proved to this data.
- Even if most of the attributes of the original data were relevant for this study, the researcher believes that other important features which can make this study more interesting were not included in the paper based file. For example, the feature which can show whether the patient with diarrhea had died or cure after treatment was not included. If this was included from the original data, it could be possible to classify and predict the mortality trend of diarrheal disease in the city.
- Data mining applications in healthcare can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry consider how data can be better captured, stored, prepared, and mined. For example, standardization of clinical vocabularies enhances the benefits of healthcare data mining applications. Therefore, creation of awareness about the advantages of data mining in the health care sector is necessary.
- To sum up, the effective use of information and technology is crucial for health care organizations to stay competitive in today's complex environment. The challenges faced when trying to make sense of large, diverse, and often complex data source are considerable. In an effort to turn information into knowledge, health care organizations should implement data mining technologies to help control costs and improve the efficacy of patient care.

References

1. A.Mekasha and A. Tesfahun (2003). Determinants of diarrheal diseases: A Community Based Study in Urban South Western Ethiopia. East African Medical Journal, 80(2), 77-78.
2. Andualem, A. and Abera,K.(2010). Assessment of the impact of latrine utilization on diarrheal diseases in the rural community of Hulet Ejju Enessie Woreda, East Gojjam Zone, Amhara Region , Ethiop. J. Health Dev. 24(2), 116.
3. Barker, R. (2011). Managing a Data Warehouse. Veritas Software Corporation, Chertsey, UK.
4. Baylis, P. (2008). Better health care with data mining. Clementine – Working with health care, Shared Medical Systems Limited,White Paper, UK.
5. Beddow, V. (2010). Sanitation Status. Addis Ababa, Ethiopia.
6. Bellaachia, A. and Guven, E. (2005). Predicting Breast Cancer Survivability Using Data Mining Techniques.Department of Computer Science,The George Washington University,Washington DC.
7. Bhargavi, P. and Jyothi, S. (2009). Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils.International Journal of Computer Science and Network Security, 9(8), 118-119.
8. Birnbaum, D. and Obenshain,M. (2004). Application of Data Mining Techniques to Healthcare Data, 25(8), 690-695.
9. Bouckaert, R., Frank,E., Hall,M., Kirkby,R., Reutemann,P., Seewald,A. and Scuse,D. (2010).A Weka Manual for Version 3-6-2.University of Waikato, New Zealand,14-21.
10. Boulle, M. (2007).Compression-Based Averaging of Selective Naive Bayes Classifiers. Journal of Machine Learning Research 8(-), 1659-1662.
11. Bresfelean,V. (2007). Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, Cavtat, Croatia.
12. Cabell, S. (2006). Data mining and Health care. Theoretical Foundations of Nursing Informatics in the School of Nursing, Troy University, Norfolk, Virginia.
13. Caelli, T. and Bischof, W. (1997). Machine Learning and Image Interpretation, New York, NY, USA: Plenum Press.
14. Ceci, M. (2005), Naive Bayesian Learning from Structural Data. A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science in the Graduate Division of the University of Bari, Italy.
15. Chye, K. and Gerald,T. (2007). Data mining applications in healthcare. Research Report.
16. Damte, S. Daniel, B. and Debela, C. (2008). Effect of Zinc supplementation in

treatment of acute diarrhea among 2-59 months children treated in Tikur Anbessa Hospital. Addis Ababa, Ethiopia, 22(2), 187-190.
17. Degruy,K.(2000). Healthcare Applications of Knowledge Discovery in Databases. Journal of Healthcare Information Management, 14(2), 59-68.
18. Delen, D. and Sirakaya, E.(2006). Determining the Efficacy of Data-mining Methods in Predicting Gaming Ballot Outcomes. Journal of Hospitality & Tourism Research, 30(3), 317-322.
19. Desalegn,S. (2009). Strange disease causing several deaths in Addis not unexpected. Addis Ababa, Ethiopia.
20. Desouza,K.(2001). Artificial intelligence for healthcare management. In Proceedings of the First International Conference on Management of Healthcare and Medical Technology, Institute for Healthcare Technology Management, Enschede, Netherlands.
21. DT, J., RG, F. and MW M. (2006). Disease and Mortality in Sub-Saharan Africa. 2nd ed., Washington (DC).
22. Eapen, G. (2004). Application of Data mining in Medical Applications. A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree Master of Applied Science in Systems Design Engineering, Waterloo, Ontario, Canada.
23. Elovici,Y.,Shabtai,A.,Moskovitch,R.,Tahan,G. and Glezer,C. (2007). Applying Machine Learning Techniques for Detection of Malicious Code in Network Traffic. Deutsche Telekom Laboratories at Ben-Gurion University, Israel.
24. El-Telbany,M.,Warda,M.and El-Borahy,M. (2006). Mining the classification Rules for Egyptian Rice Diseases. The international Arab journal of information Technology,3(4) ,303-305
25. Federal Ministry of Health (FMoH) (2005). National Strategy for Child Survival in Ethiopia. Addis Ababa, Ethiopia.
26. Felkey, G. Liang,H.and P. Krueger,K. (2003). Data Mining for the Health System Pharmacist. Journal of Hospital Pharmacy, 38(9), 849.
27. Ferrari, D. (2005). Mining Housekeeping genes with a Naive Bayes classifier. Master of Science School of Informatics, University of Edinburgh.
Five Children in Rural Upper Egypt. Journal of tropical pediatrics, 46,283-284.
28. G.Mitikie (2001). Prevalence of acute and persistent diarrhea in North Gonder Zone, Ethiopia. East Africa Medical Journal, 78(8), 433-436.
29. Gerritsen, R. (1999). Assessing Loan Risks:A Data Mining. Case Study.
30. Gerver,H. and Barrett, J. (2006). Data mining–Driven ROI: Healthcare cost Management. Benefits and compensation Digest, 43(6), 1-4.
31. Girma, R. Wondwossen, B., Bishaw, D. and Tefera, B. (2008). Environmental Determinants of Diarrhea among Under-Five Children. Ethiop. J. Health Sci., 18(2), 39-41.
32. Glasgow,J.and Ng,R.(1999). Data Mining and Knowledge Discovery in Molecular Databases.

33. Glover, S., Rivers,P., Asoh,D., Piper,C. and Murph,K. (2010). Data mining for health executive decision support: an imperative with a daunting future! Health Services Management Research,23(1),42-44.
34. Hajizadeh,E., Ardakani,H. and Shahrabi,J. (2010). Application of data mining techniques in stock markets: A survey. Journal of Economics and International Finance, 2(7), 109-113.
35. Hamer,D.,Simon,F.,Thea,D. and Keusch,G. (1998).Childhood Diarrhea in Sub-Saharan Africa. Child health research project special report, 2(1), 3-11.
36. Hamilton,H.,Gurak,E.,Findlater,L. and Olive,W.(2011). Overview of Decision Trees**,** Rudjer Boskovic Institute.
37. Hand, D., Mannila, H., and Smyth, P. (2001). Principles of Data Mining. MIT Press, Cambridge.
38. International Federation of Red Cross and Red Crescent societies, Ethiopia (ERCS) (2009). Acute Watery Diarrhea. DREF operation final report.
39. International Federation of Red Cross and Red Crescent societies, Ethiopia (ERCS) (2010). Acute Watery Diarrhea. DREF operation final report.
40. Kaur, H. and Wasan, S. (2006). Empirical Study on Applications of Data miningtechniques in Healthcare. Journal of Computer Science, 2(2), 194-198.
41. Kerdprasop, N. and Kerdprasop, K. (2009). Knowledge Induction from Medical Databases with Higher-Order Programming, 6(10), 1719-1723.
42. Keusch, G., Fontaine, O., Bhargava, A., Boschi, C., A. Bhutta, Z., Gotuzzo, E., Rivera, J., Chow, J., A. Shahid ,S., and Laxminarayan, R. (1993). Diarrheal Diseases.
43. Koh, H. and Tan, G. (2005). Data Mining Applications in Healthcare. Journal of healthcare information management, 19(2), 64.
44. Kraft,M., Desouza,K. and Androwich,I. (2002). Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population.
45. Kudyba, S. and Gregorio, T. (2010). Identifying factors that impact patient length of stay metrics for healthcare. Journal of Health Informatics, 16(4), 236-239.
46. Kumar, V. and Nisha, R. (2011). Knowledge discovery from database using an integration of clustering and classification. (IJACSA) International Journal of Advanced Computer Science and Applications,2(3),29-32.
47. Magendram, A. (2007). Classification System for Heart Disease Using Bayesian Classifier. Thesis submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in the Faculty of Computer Science and Information Technology,Putra University, Malaysia.
48. Mariscal, G., Marban, O. and Fernandez,C. (2010). A survey of data mining and knowledge discovery process models and methodologies.The Knowledge Engineering Review, 25(2), 137-149.
49. Mirgissa, K. and Fikau, A. (2000).Ethnographic study of diarrheal diseases among under five children in Mana District,Jimma Zone,South West Ethiopa. Ethiopian journal of health development,14(1),77-83

50. Nasereddin, H. (2009). Stream Data Mining. International Journal of Web Applications, 1(4) ,183-188.
51. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (2003),Diarrhea,NIH Publication.
52. Obenshain, M. (2004), Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 25(8), 690-690.
53. Palaniappan, S. and Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International Journal of Computer Science and Network Security, 8(8), 343-345.
54. Patil,S. and Kumaraswamy,Y. (2009) Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network. European Journal of Scientific Research, 31(4), 643-650.
55. Pedarla, P. (2004). E-Intelligence form design and Data Preprocessing in Health Care. The University of Waterloo in fulfillment of the thesis requirement for the degree of Master of Applied Science in Systems Design Engineering, Waterloo, Ontario, Canada.
56. Pop, I. (2006). An approach of the Naive Bayes classifier for the document classification. General Mathematics, 14(4),136.
57. Ranjan, J., Nagar,R., and Pradesh,U. (2007). Applications Of Data Mining Techniques In Pharmaceutical Industry. Journal of Theoretical and Applied Information Technology, 3(4), 61-65.
58. Rañoa, C. (1983). Cost Reduction in the Treatment of Diarrheal Diseases by Oral Therapy.
59. Rokach, L. & Maimon, O. (2005). Top-down induction of decision trees classifiers – a survey: Applications and Reviews, IEEE Transactions, 35(4), 476-487.
60. Rygielski,C.,Wang,J. and Yen,D.(2002). Data mining techniques for customer relationship management. Technology in Society, Taiwan.
61. Safran, C., Mery, H., W. Edward, L., Markel, S., Tang, P., and Detmer D. (2007). Toward a National Framework for the Secondary Use of Health Data. Journal of the American Medical Informatics Association, 14(1), 1-5.
62. Sazawal,S.,RE,B. and MK,B. (2011). Zinc Supplementation Reduces the Incidence of persistent diarrhea and dysentery among low socio-economic children. Journal of nutrition, India.
63. Scioinspire Corporation (SIC) (2009). Predictive modeling basics.
64. Shailja (2009). Classifying Web Services with and without Association Rules. Thesis submitted in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science & Engineering, Thapar University, Patiala.
65. Shearer, C. (2000).The CRISP-DM Model: The New Blueprint for Data Mining. Journal of data warehousing, 5 (4), 13-18.
66. Shegaw, A. (2002). Application of data mining technology to predict child mortality patterns: the case of Butajira Rural Health Project (BRHP). A thesis

submitted in partial fulfillment of the requirement for the Degree of Masters of Science in Information Science, Addis ababa University.
67. Silver,M., Su,H.and Dolins,S.(2001). Case Study: How to Apply Data Mining Techniques in a Healthcare Data Warehouse. Journal of Healthcare Information Management,15, (2), 155-157.
68. Singh, P. (2009). Comparing the Effectiveness of Machine Learning Algorithms for Defect Prediction. International Journal of Information Technology and Knowledge Management, 2(2), 482-483.
69. Soman, T. and Bobbie, P. (2005). Classification of Arrhythmia Using Machine Learning Techniques. Southern Polytechnic State University (SPSU), S. Marietta Parkway, Marietta, USA.
70. Srinivas, K.,Rani,B. and Govrdhan,A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 2(2),250-254.
71. Stolba, N. and Tjoa, A. (2005).The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making. World Academy of Science, Engineering and Technology, Vienna University of Technology, Austria.
72. Subbalakshmi, G., Ramesh,K. and Rao, M.(2011). Decision Support in Heart Disease Prediction System usingNaive Bayes, Indian Journal of Computer Science and Engineering (IJCSE) 2 (2),172-174.
73. Trybula, J. 1997. Data Mining and Knowledge Discovery: Annual review of Information Science and Technology (ARIST), 32, 197-216.
74. Two Crows Corporation (TCC) (1999). Introduction to Data Mining and Knowledge Discovery.
75. Two Crows Corporation (TCC) (2005). Introduction to Data Mining and Knowledge Discovery.
76. U.S. Agency for International Development (USAID) (2005). Guidelines for New Diarrhea Treatment Protocols for Community-Based Healthcare Workers. The most project.
77. United Nations Children's Fund (UNICEF) and World Health Organization (WHO) (2009). Diarrhea: Why children are still dying and what can be done? WHO Library Cataloging-in-Publication Data.
78. Untwal, L. (2008). Data Mining: A Handy Tool for Pharmaceutical Industry.
79. Wasan,K., Bhatnagar, V. and Kaur H. (2006).The Impact of Data Mining Techniques on Medical Diagnostics. Data Science Journal, 5(19), 119-124.
80. Witten,I.,Frank,E.,Trigg,L.,Hall,M.,Holmes,G. and Cunningham,S. (2000). Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers.
81. World Health Organization (WHO) (1995). Reducing Risks, Promoting Healthy Life. The World Health Report.
82. World Health Organization (WHO) (2000).The global burden of diarrheal disease, as estimated from studies published between 1992 and 2000.

83. World Health Organization (WHO) (2004). Diarrheal disease. The World Health Report.
84. World Health Organization (WHO) (2005).The treatment of diarrhea, a Manual for Physicians and Other Senior Health Workers.
85. World Health Organization (WHO) (2005. Diarrheal disease. The World Health Report.
86. World Health Organization (WHO) (2007). Emergency Humanitarian Action (EHA) / Ethiopia Programme. Country Offices Monthly Report, Ethiopia.
87. World Health Organization (WHO) (2008).Emergency and Humanitarian Action (EHA) Weekly Update/week 3/Ethiopia.
88. World Health Organization (WHO) (2011).The global burden of diarrheal disease. The World Health Report.
89. Xhemali, D., J. Hinde,C. and G. Stone, R. (2009). Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. IJCSI International Journal of Computer Science Issues, 4(1),17-21.
90. Yassin, K. (2000).Morbidity and Risk factors of Diarrheal diseases among Under Five Children in Rural Upper Egypt. Journal of tropical pediatrics, 46(3),283-284.

Appendices

Appendix A:

Selected attributes with their description and Attribute type

Attribute	Description	Attribute Type
Year	Patient's year of treatment for diarrhea disease	Nominal
Age	Patients age at the time of treatment	Numeric
Sex	Sex of patients who were provided treatment	Nominal
Weight	The weight of the patient at a time of treatment	Numeric
Visits	Duration of time the patient come to the center(N=New,R=Repeat)	Nominal
Nutritional Status	The percentage of nutrition of the patient with diarrhea at the time of treatment	Nominal
Type of diarrhea	The clinical type of diarrhea observed on the patient at a time of treatment	Nominal
Degree of dehydration	The amount of fluid lost from patients' body as a result of diarrhea before treatment	Nominal
Treatment	The type of treatment provided for the patient with diarrhea	Nominal

Appendix B:

Target Class Attributes and their Values

Attribute Class	Value Lable
Treatment	{ORS,ORS and RL,RL}
Type of diarrhea	{WATERY,BLOODY,PERSISTENT}

Appendix C:

Results of some trials before and after parameter settings modification using J48 decision tree classifier for target class 'Treatment'.

Trials	'MinNumobj'	'ConfidenceFactor'	Size of A tree	Number of leaves	Accuracy
1	Default (2)	Default(0.25)	636	393	91%
2	25	0.025	212	139	88%
3	25	0.0025	201	133	87%
4	40	0.0025	168	115	86%
5	55	0.0025	145	103	85%

Appendix D:

Results of some trials before and after parameter settings modification using J48 decision tree classifier for target class 'Type of Diarrhea'.

Trials	'MinNumobje'	'ConfidenceFactor'	Size of A tree	Number of leaves	Accuracy
1	Default (2)	Default(0.25)	1163	747	91%
2	5	0.025	645	415	89%
3	15	0.025	488	328	87%
4	25	0.025	366	250	86%
5	25	0.0025	325	224	85%