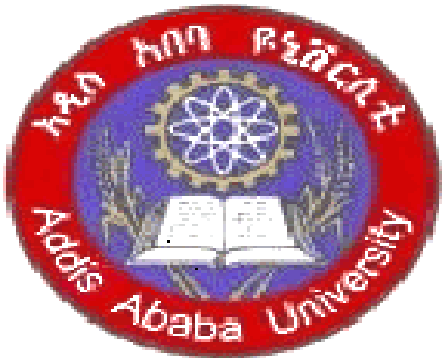


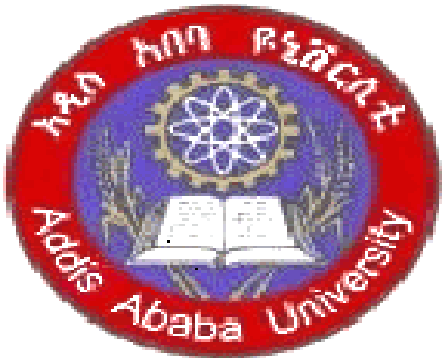
ADDIS ABABA UNIVERSITY
SCHOOL OF PUBLIC HEALTH AND SCHOOL OF INFORMATION SCIENCE

MASTER OF SCIENCE IN HEALTH INFORMATICS

Application of Data Mining Techniques to Discover Cause of Under-five Children Admission to Pediatric Ward:
The case of Nigist Eleni Mohammed Memorial Zonal Hospital

By
Temesgen Dileba

June, 2012
Addis Ababa



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

MASTER OF SCIENCE IN HEALTH INFORMATICS

Application of Data Mining Techniques to Discover Cause of Under-five Children Admission to Pediatric Ward: The case of Nigist Eleni Mohammed Memorial Zonal Hospital

By
Temesgen Dileba

A thesis presented 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.

June, 2012
Addis Ababa

DECLARATION

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

Name of candidate_____

Date_____ Signature_____

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

Name of advisor_____

Date_____ Signature_____

Name of advisor_____

Date_____ Signature_____

Approval of the examining board

Name _____

Date_____ Signature_____

Name _____

Date_____ Signature_____

Name _____

Date_____ Signature_____

Acknowledgement

Before all, I would like to extend my gratitude and limitless respect to my almighty God who gave me strength throughout the quest in this thesis. Next my immense appreciation and honest thanks go to my advisors Dr. Dereje Teferi (Ass. Prof.) and Dr. Assefa Seme (MD, MPH) for their unreserved comments, guidance, encouragement and constructive suggestions starting from title selection to the last point of this thesis.

I would also like to extend my sincere thanks to Dr. Ayano Dr. Adane, S/r Selam Tessema, Ato Genene Teshome, Ato Yeshano, W/ro Ayehubirihan and all staff members of pediatric ward and under-five OPD of NEMM hospital for their sharing experience, cooperation, support and encouragements during data collection and output interpretations and others who played great role in reaching me to this position. My deepest thanks also refer to all staff members (especially, ATO Dawit Biwota) and to my classmates (especially Minale Tefera) for their motivation and follow up all the way through the compilation of my study.

My heartfelt thanks go to my brothers and relatives especially, Teketay, Alemayehu, Hanka and my friends Mekonin Bogale, and Yosef who gave me all rounded support in the journey of this thesis. Next, I would like to thank School of Information Science, Department of Health Informatics and Public Health and Addis Ababa University for the financial support and overall facilitation of the thesis progress.

Finally, special thanks go to my wife W/o Abaynesh Kafe for her devotion in handling all family related issues on behalf of me for the past two years.

Table of Contents

List of tables

List of figures

Acronyms

AGE-Acute Gastroenteritis

ANN-Artificial Neural Network

AWFA-Appropriate Weight for Age

CMR-Child Mortality Rate

CRISP-DM-CRoss-Industry Standard Process for DM

EDA-Exploratory Data Analysis

HMIS-Health Management Information System

LWFA-Low Weight for Age

FMOH- Federal Ministry Of Health

MUAC-Mid-Upper Arm Circumference

NEMM-Nigist Eleni Mohammed Memorial

NMR-Neonatal Mortality Rate

NGO-Non-Governmental Organization

OPD- Out Patient Department

RVI-Retroviral Infection

SAM-Severe Acute Malnutrition

SMOTE- Synthetic Minority Over-sampling TEchnique

SNNPR-South Nation Nationalities Peoples Region

VWFA-Very low Weight for Age

WEKA- Waikato Environment for Knowledge Analysis

Abstract

Background: - Health care system is potential area to apply and take the advantage of data mining. Higher priority is given for the prevention and control of preventable disease at home or community level. However, for seriously ill children admissions should be facilitated in order to save the life of the child.

Objectives: - The objective of this study is to apply data mining techniques on under five children dataset in developing a model that support the discovery of the causes for under-five children admission to pediatric ward.

Methodology: - Cross industry standard process for data mining process model was applied. Major processes covered were business understanding, data understanding, data preprocessing, modeling and evaluation. Decision tree and artificial neural network algorithms were tested for classification tasks in Waikato Environment for Knowledge Analysis. Exploratory data analysis techniques, graphs and tabular formats for visualization and accuracy, true positive rate, false positive rate, receiver operating characteristic and the idea of experts were used for evaluation of the model. The dataset used was records in integrated registration log book in under-five outpatient department.

Result: - The decision tree algorithm J48 has higher accuracy (94.77%), weighted true positive rate (94.7%), weighted false positive rate (5.3%), weighted receiver operating characteristics (0.99) and performs much faster than multilayer perceptron. According to interesting rules in J48 presenting complaint of not taking any food, fluid or breast feeding (98.32%), low weight for age without sunken eyes (92.31%) and very low weight for age but not in association with restless or irritable (98.33%) are among the cause of under-five children admission to pediatric ward without any consideration of health information management system admission disease classification.

Conclusion: - In conclusion, encouraging results are obtained in classification tasks, data mining technique is applicable on pediatric dataset in developing a model that support the discovery of the causes of under-five children admission to pediatric ward. The outcome of this study serves primarily users in the domain area, decision makers and planners.

CHAPTER ONE

INTRODUCTION

1.1 Background

There are tremendous volumes of data filling our computers, networks, and lives. Government agencies, scientific institutions, and businesses have all dedicated enormous resources to collecting and storing data. In reality, only a small amount of these data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analyzed effectively. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world (1).

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration (2).

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from this world (such as supermarket transaction data, government statistics, etc) to the more exotic (such as molecular databases, and medical records). Interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database (3).

Data mining is an iterative process and search for new, valuable, and nontrivial information in large volumes of data and needs a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers (1).

Clinicians evaluate a patient's condition over time. The analysis of large quantities of time-stamped data will provide doctors with important information regarding the progress of the disease. Although the use of temporal-reasoning methods requires an intensive knowledge-

acquisitions effort, data mining has been used in many successful medical applications, including data validation in intensive care, the monitoring of children's growth, analysis of diabetic patient's data, the monitoring of heart-transplant patients & intelligent anesthesia monitoring (1).

A child born in a developing country is over 13 times more likely to die within the first five years of life as compared to economically advanced countries. Sub-Saharan Africa accounts for about half of these deaths in the developing world. Surprisingly the causes of illnesses and admissions to hospitals that utilize the scarce resource in the region are diseases that can be easily prevented. Child illnesses and deaths are higher for children from rural and poor families and whose mother lack basic education. An Ethiopian child is 30 times more likely to die by his or her fifth birth day than a child in Western Europe (4).

Hospitals and health centers are primarily medical care centers where huge data is collected regarding the patients and clients in daily bases. Unless this data is stored and processed to provide appropriate information, it has no value. Appropriately processed Information is a good source for health managers at different level of organizations in order to make informed decision. But in our case particularly in Ethiopia, the data is there we are starving from knowledge due to lack of capacity in turning data into information and then to use it as an input for decision making. Therefore, the ability to use these data to extract useful information for quality health care is a crucial issue. As indicated above, data mining is one of the solutions to analyze large amount of data and thus into information and knowledge.

Easily preventable diseases are the major causes of illness in developing countries. The burden is highest in sub-Saharan Africa. Child mortality rate is also high in Ethiopia (104/1000) (5). The major causes of illnesses are preventable diseases that can be controlled easily. Hence, it is still an urgent area that needs the attention of government, NGO’s, research centers, MOH, and other stakeholders to come through updated and new strategies, solutions and researches.

1.2 Statement of the Problem

When the issue of growth and development of a country is raised, the indicators of child illness and deaths have a prominent place. Ethiopia is among the highest country in child illnesses and deaths in the world. As indicated in health and health related indicators of 2011, neonatal mortality rate is 36/1000 live birth and infant mortality rate is 67/1000 live births (5).

Pneumonia, diarrhea, malaria and measles that easily prevented through simple improvements in basic health services and interventions, are the leading cause of child illnesses and deaths (6).

Child health focuses on diseases that are causes of illnesses and death for the majority of children. Illnesses leading to hospitalization of children are primarily due to easily preventable diseases such as acute respiratory infections, diarrheal disease, malaria, measles and malnutrition. Strengthening preventive activities related to these diseases make the health service complete, and integrate and improve the quality of health care (7).

Though, there has been encouraging progress in recent years towards improving the health of children in Ethiopia, childhood illness is still a pressing problem. Study conducted on admission to the pediatric emergency ward in Tikur Anbesa hospital in Addis Ababa revealed that major causes of hospital admissions are pneumonia, meningitis, sepsis, and gastroenteritis which are preventable and treatable with available medical technology. Severe pneumonia accounted for 38.3% of the total admissions (8).

Child illnesses vary across the regions of Ethiopia. A study from South Nations Nationalities and Peoples Region has shown that children less than 5 years of age had between 6 and 12 episodes of illness per year. Hospital admissions show that acute respiratory infections, malaria, diarrhea and malnutrition are major causes of illness in this region. Malnutrition, particularly in combination with acute respiratory illness, malaria or measles is another important cause of illness in children (9).

Hospital admissions provide some measures of prevalence and severity of childhood illness. Admission rates and their cause reflect socioeconomic circumstances, the level of utilization of primary health care services and the health care seeking behavior of community. However, because of delayed health seeking behavior of parents and other care providers, easily preventable diseases may be complicated and even may lead to death of children.

Higher priority is given for the prevention and control of preventable diseases at home and community level. However, for seriously ill children admissions should be facilitated in order to save the life of the child. This assists in managing the cases and facilitates the conditions for the next admission in order to utilize resources appropriately.

There is a rapidly widening gap between data-collection and data-organization capabilities and the ability to analyze the data. Whether the context is business, medicine, science, or

government, the datasets themselves, in their raw form, are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. The root of the problem is that the data size and dimensionality are too large for manual analysis and interpretation, or even for some semiautomatic computer-based analyses (1).

Large data is collected in NEMM hospital about patients in different departments in daily bases. Under-five OPD and pediatric ward are among departments to manage under-five OPD visit and admit children below the age of 15years in NEMM hospital. Different attributes of each child is recorded in under-five OPD. However, the data is yet not processed, analyzed and utilized for planning and to make informed decision. Usually data is collected for the purpose of reporting and then thrown in the record room. Thus, huge information remains hidden.

The underlying research problem that necessitated this research is the fact that, although large amount of data is available, they are not using it in a way that supports their objectives. Decision making bodies are not using this data for making informed decision. Thus, the data remains unutilized to the problems faced by the society due to lack of research in deploying appropriate data analysis and mining tools.

Therefore, this research was applied to discover the cause of under-five children admission by using data mining techniques on large data in pediatric dataset of NEMM hospital. The findings will be taken as one input and to come up with new knowledge in the area that enhance the decision making of managers at different level of health sector. This study attempts to answer research question of this study, from the data available in under- five OPD, in that:

- Preparing the data set for mining under-five OPD visit data
- Extracting hidden knowledge that is available in under- five children admission
- Exploring the common illnesses that are contributing to under-five children admission

1.3 The rationale of the study

The model built by using different attributes recorded in under-five OPD visit data set classify the outcome of under five OPD visit (either admit or not admit to pediatric ward). These assist physicians and other health care workers in managing children visiting under five OPD. Prediction of admissions also helps patients to get appropriate treatment at timely so that, an unnecessary delay in case management is also resolved. Furthermore it facilitates the overall

communication line among under- five OPD and pediatric ward regarding to beds, critical case management, and appropriate utilization of other resources.

1.4 OBJECTIVES

1.4.1 General Objective

The general objective of this research is to apply data mining techniques in developing a predictive model that support the discovery of the causes of under five children admission at Nigist Eleni Mohammed Memorial hospital.

1.4.2 Specific Objectives

In order to achieve the general objective indicated above, the research has the following specific objectives:

- Assess the requirements, and overall working procedures in the hospital
- Extract the dataset required for analysis from under-five OPD and pediatric ward of NEMM hospital
- Prepare the data for analysis and model building by cleaning, extracting and transforming into a format suitable for the selected data mining algorithm
- Select an appropriate data mining tools and techniques to be used for data mining
- Apply classification algorithms to classify, build, train, test and compare the classification models that classify instances of children

1.5 Significance of the study

There was no data mining technique that has been applied to discover cause of under-five children admission and length of stay in pediatric ward elsewhere in Ethiopia. Lack of adequate research in the application of data mining technique to discover cause of under-five children admission and length of stay in pediatric ward justifies a new research that can handle the large data available in pediatric ward in Nigist Eleni Mohammed Memorial hospital.

The result from this study applied in various areas. Firstly, the findings provide an input for physicians and health care workers to have reliable information on a child's requirement for admission based on preliminary investigation of the child's sign and symptom at Nigist Eleni Mohammed Memorial hospital. For ministry of health and non-governmental organizations who

are involved in research and planning, the result lead or facilitate a direction for informed decision making regarding to the under-five child admission.

Secondly, the result of the research used by policy makers who are expected to make decision about under-five children disease prevention and control programs and sat policies based on research findings. Thirdly, this research fills the gaps around the health sectors in using data mining techniques. Researchers who are involved on child health studies may use the results and recommendations to search the gap in order to fulfill uncovered areas.

1.6 Scope of the study

The scope of this research is limited to apply data mining techniques to discover causes of under-five children's admission in order to build predictive model for outcomes of under-five OPD visit. The result is described by using classification analysis function of data mining to under five children's dataset. The study did not include under-five children visits from other departments. The result of investigation is primarily applicable in under-five outpatient department of Nigist Eleni Mohammed Memorial hospital and other similar settings.

1.7 Organization of the Thesis

The thesis is organized into five chapters. Chapter one is an introductory chapter, which is regarding to background, statement of the problem with research questions, research objectives the scope and research organization. Chapter two provides overview about data mining, models, algorithms and tools and also describes child health definitions, concepts and variables in a brief way. In chapter three details of methodology followed by the researcher is discussed. In chapter four the experimentation and discussion of the result are included where overview about experimentation, data selection process, and clustering sub phase, classification sub-phase, and discussion of the result are presented. Chapter five contains the conclusion and recommendation part of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Data Mining

In the era of information technology, one interesting opportunity is using information for its intended purpose. The daily source of information (aggregation of data that makes decision making easier) is data (unorganized and unprocessed facts; a set of discrete facts about events). These data are available in various forms such as primitive file processing systems to more sophisticated databases, data warehouses, World Wide Web and data streams (2).

Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities such as data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and advanced data analysis (involving data warehousing and data mining). The early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, advanced data analysis has naturally become the next target (2).

Huge volumes of data can be accumulated through time. The effective and efficient analysis of these data and changing it into information and knowledge is a crucial issue. Deficiency on these analyses may create a data rich but information poor situation. In the information society, advance in ICT have made data easy to use and cheap to store and exchange. Database across the world contain data that exist in digitized text document, video and audio files and financial transactions (18). Even if we are surrounded by huge data, we are facing difficulty in getting information and /or knowledge. The available data, in itself is not enough for improving the work and to have meaning and relationship. We need to able to transform the raw data into information which is useful for taking important business decisions. According to this, it requires an especial tool for efficient and effective analysis of such data. Data mining is the technology needed for extracting or “mining” knowledge from large amounts of data (2).

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. It is a structured way of finding out what potential information it contains and how it applies to solving the business problem (1). It is the extraction of implicit, previously unknown, and potentially useful information from data (10).

When we see the evolution of data mining, it does not mean that data mining is different from other disciplines such as statistics. In statistics, researchers frequently deal with the problem of finding the smallest data size that gives sufficiently confident estimates. Data mining deals with the opposite problem, that, data size is large and we are interested in building a data model that is small but still describes the data well (11, 12).

The two primary goals of data mining are prediction and description (1). Prediction involves using some variables or fields in the dataset to predict unknown or future values of other variables of interest. It produces the model of the system described by the given dataset. The goal is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. It produces new, nontrivial information based on the available dataset. The goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets (1).

To achieve the goals of prediction and description the following data-mining techniques are used*:* Classification is discovery of a predictive learning function that classifies a data item into one of several predefined classes. Regression is discovery of a predictive learning function, which maps a data item to a real-value prediction variable. Clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data. Summarization is an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data. Dependency Modeling is finding a local model that describes significant dependencies between variables or between the values of a feature in a dataset or in a part of a dataset. Change and deviation detection is discovering the most significant changes in the dataset. Association rules are also one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems (1, 12).

Traditional statistical studies use past information to determine a future state of a system whereas data mining studies use past information to construct patterns based not only in the input data, but also on the logical consequence of those data. It also contains a vital element missing in statistical analysis: the ability to provide an orderly expression of what might be in the future, compared to what was in the past. Compared to traditional statistical studies the findings of data mining, patterns and classifications look forward and even predict future. However, the most important challenges and problems of data mining are noise data, difficult training set, and the nature of the database and the size of the database. As it is relatively new field, extracting useful information from the data may be complicated and even difficult challenge (13).

As a field of study, scholars are explaining theories and definition of the data mining field. They have come up with numerous definition of data mining. Some of them are summarized as follows:

Data mining refers to extracting or "mining" knowledge from large amounts of data (2). It is a term that covers a broad range of techniques being used in a variety of industries. Due to increased competition for profits and market share in the marketing arena, data mining has become an essential practice for maintaining a competitive edge in every phase of the customer lifecycle (14).

Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (3). When it is summed up, all scholars define data mining as; it is a technology that uses various techniques to discover hidden knowledge from heterogeneous and distributed historical data stored in large databases, warehouses and other massive information repositories.

2.2 Why Data Mining?

When we see the evolution of data mining, it does not mean that data mining is completely different from other disciples and field of studies like statistics and data warehouse. Even if data mining has come to existence in response to technological advances in many diverse disciplines, it is an "umbrella" term coined for the purpose of making sense of data. Because other fields like statistics also deals with data they further argued that data mining is a data driven approach, as opposed to model driven. In statistics, researchers frequently deal with the problem of finding the smallest data size that gives sufficiently confident estimate. Data mining deals with the opposite

problem, namely, data size is large and we are interested in building a data model that is small but still describes the data well (12, 15).

Data warehouse provides clean and integrated data for fruitful mining. Due to this there is some real benefit if the data is already part of data warehouse. It is better to use data mining on data stored in data warehouses because it can provide powerful tools (e.g., association, classification, and clustering and trend analysis) for analysis (22).

2.3 Data Mining and Knowledge Discovery

Knowledge discovery in database (KDD) is defined as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns of data. One of the crucial steps in KDD is data mining; however, they are often used as synonyms (23, 24).

As indicated in Figure 2 .1 below KDD is defined as the whole process involving: data selection, data pre-processing: cleaning, data transformation, mining, result evaluation and visualization. On the other hand, data mining refer to the modeling step using the various techniques to extract useful information/pattern from the data. Hence, KDD is the process of finding useful information and patterns in data and data mining is the use of algorithms to extract hidden patterns & knowledge in data.

2.4 Data mining Models and Methods

Data mining could be applicable to any kind of information repository. This includes relational database systems, data warehouse, transactional database, flat files and World Wide Web in order to develop models. The challenges and techniques of mining may differ for each of the repository systems (2).

A model is an abstract representation of a real-world process. For example, $Y = 3X + 2$ is a very simple model of how the variable Y might relate to the variable X (3). Two primary goals of data mining in practice are prediction and description in order to develop a model that predict or describe the variables (1, 24).

Knowledge discovery applies a set of appropriate algorithms and mechanisms to extract and present the knowledge from a given dataset, i.e. identifying valid, novel, potentially useful, and ultimately understandable patterns in the dataset (23). Data mining is the analysis of large observational dataset to find unsupervised relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (3).

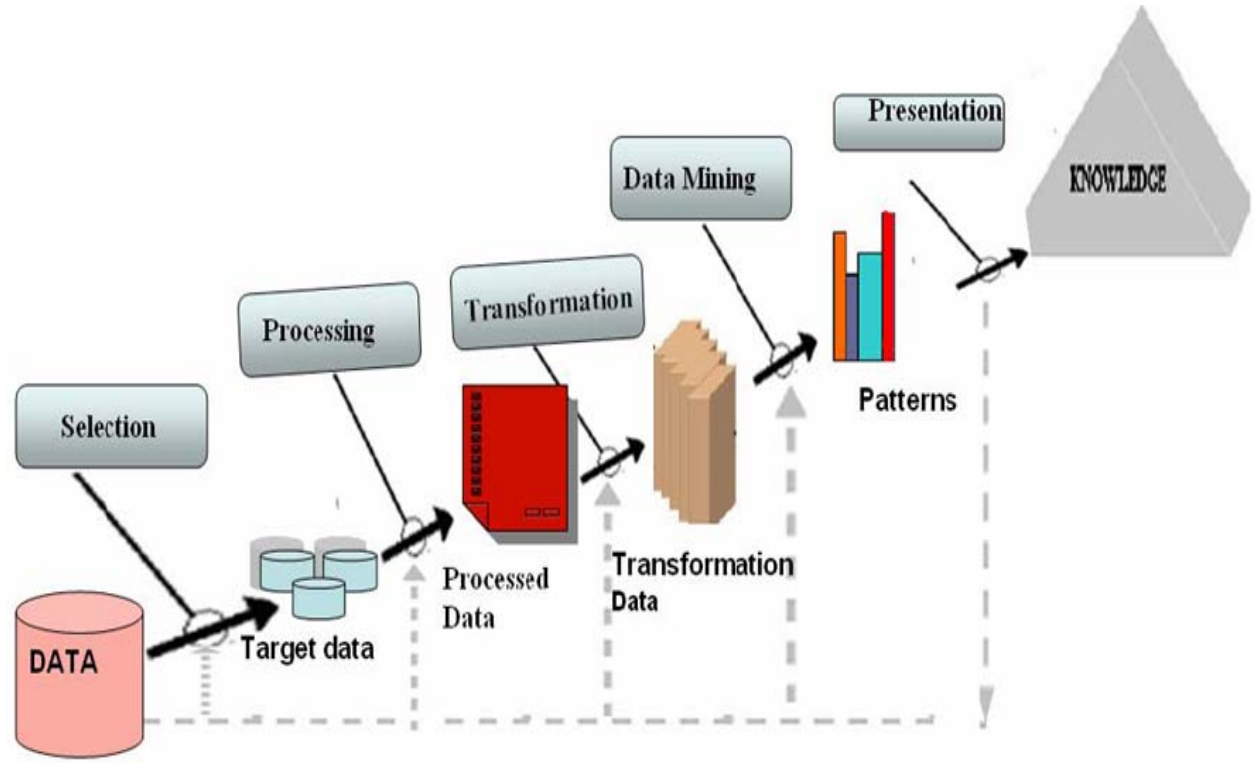


Figure: 2.1 Data mining in KDD process model

2.5 Data Mining Algorithms and Tools

In the following section data mining algorithms that are used to achieve the predictive goals of data mining are reviewed.

2.5.1 Classification Algorithm

Classification of a collection consists of dividing the items that make up the collection into categories or classes. The goal of predictive classification is to accurately predict the target class for each new record that is not in the historical data. A classification task begins with build data (also known as training data) for which the target values (or class assignments) are known. Different classification algorithms use different techniques for finding relations between the predictor attributes' values and the target attribute's values in the build data. These relations are summarized in a model; the model can then be applied to new cases with unknown target values to predict target values. A classification model can also be applied to data that was held aside from the training data to compare the predictions to the known target values; such data is also known as test data or evaluation data. The comparison technique is called testing a model, which

measures the model's predictive accuracy. The application of a classification model to new data is called applying the model, and the data is called apply data or scoring data (25).

Some well-known classification algorithms are Bayesian, decision trees, neural network, K-nearest neighbor classifiers and genetic algorithms.

2.5.1.1 Decision Tree

Decision tree is a structure that can be used to divide a large collection of records into successively smaller sets of records by applying a sequence of decision rules (26). It is a supervised learning method that constructs decision trees from a set of input-output samples. A typical decision-tree learning system adopts a top-down strategy that searches for solution in a part of the search space (1).

Decision tree consists of nodes and branches connecting the nodes. The nodes located at the bottom of the tree are called leaves and indicate classes. The top node in the tree, called the root, containing examples that are to be divided into classes. All nodes except the leaves are called decision nodes, since they specify decision to be performed at this node based on a single feature. Each decision node has a number of children nodes, equal to the number of values that a given feature assumes (15).

Tree induction: decision trees are used to predict and/or classify. There are two phases, the training and implementation. During the training phase, the data set is partitioned iteratively. During each pass (i.e. iteration), the data set is split on that feature (or attribute) that produces the most effective classification. Only those factors most significant to the partitioning are used. The implementation phase then produces decision rules which are equivalent to the partitions (or branching) created during the training phase. These rules are used to generate new information when presented with novel situations (10, 15).

Decision tree attribute selection: Selecting the most discriminatory (significant) to feature should be done to solve the problems. Attribute selection is normally done by searching the space of attribute subsets, and evaluating each one. This can be achieved by using variety of techniques such as information gain attributes evaluation and gain ration attribute evaluations (10, 15). At each node, available attributes are evaluated on the basis of separating the classes of the training examples. An evaluation (goodness) function is used. For this purpose information gain (ID3/C4.5), information gain ratio, and gini index (CART) are used (2). Decision trees become

incomprehensible when their size grows; to avoid the problem of over fitting decision trees are pruned down in such a way that there is no significant loss of classification accuracy. The pruning process can be depending on when the pruning occurs during the growth process of the tree (2).

In pre-pruning, the growth of the tree stops when it is determined that no attribute will significantly increase the information gain in the process of classifying the data. While in post pruning, involves already-constructed trees. Complexity of the tree is resulted in observed loss in classification accuracy hence in order to make a good decision much of tree branches should be eliminated.

2.5.1.2 Artificial Neural Network (ANN)

Artificial neural network is an abstract computational model of the human brain. It has the ability to learn from experiential knowledge expressed through inter unit connection strengths, and can make such knowledge available for use. It has the following capabilities; a typical neural network is composed of a potentially large number of neurons arranged in three different conceptual layers: an input layer representing the input variables, one or more hidden layers, and an output layer representing the output variables (1).

A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response nodes (22).

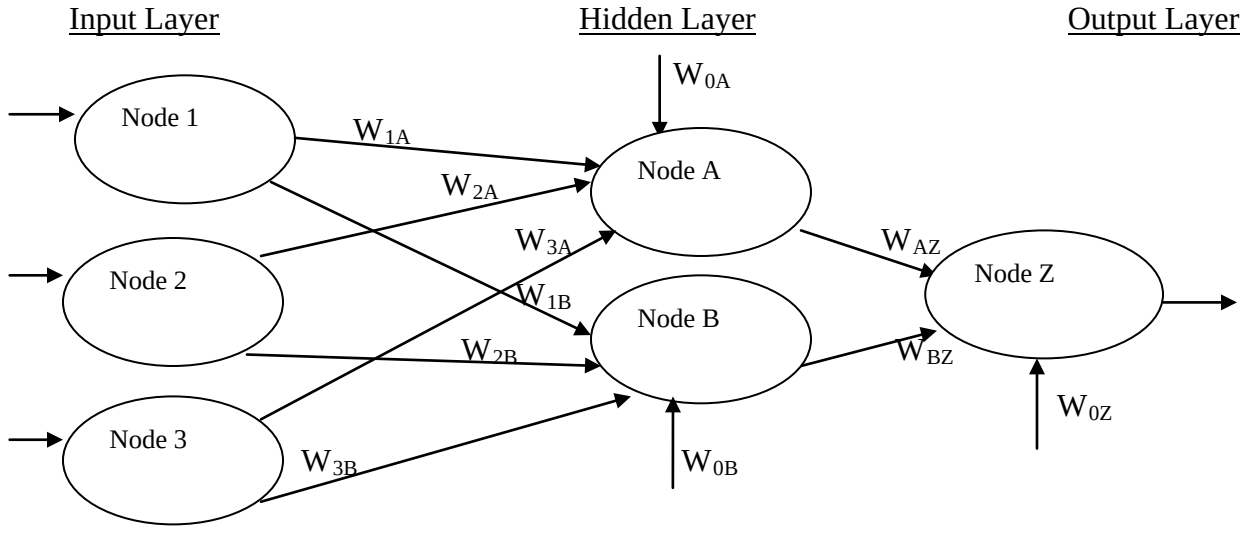


Figure: 2.2 Simple neural networks. Source: An introduction to data mining (27).

As indicated in figure 2.3 W_{1A}, W_{2B} ….W_{3B} represents weights of the nodes. Every node in a given layer is connected to every node in the next layer, although not to other nodes in the same layer. Each connection between nodes has a weight (e.g. W_{1A}) associated with it. At initialization, these weights are randomly assigned to values between zero and one. The number of input nodes usually depends on the number and type of attributes in the data set. The numbers of hidden layers, and the number of nodes in each hidden layer, are both configured by the user.

On computation of the weight value and methods used, the connection weights (W's) are the unknown parameters which are estimated by a training method (back propagation). First a combination function (usually summation) produces a linear combination of the inputs and the connection weights into a single scalar value. This will give the net for a given node. Once the net value for each node is known; it will be used as an input to the activation function (usually sigmoid function) which is used to generate the output signal from the weighted average of inputs. For the total output node Z, net $_z$ can be calculated (22).

The sigmoid function value of Net is output value of the neural network for the first pass through the network and represents the value predicted for the target variables for the first observation.

Neural network structure: there are three main types of ANN structures-single layer feed forward network, multi-layer feed forward network and recurrent networks. The most common type of single layer feed forward network is the perceptron. Other types of single layer networks are based on the perceptron model (27). Inputs to the perceptron are individually weighed and then summed. The perceptron computes the output as a function F of the sum. The activation function, F is needed to introduce nonlinearities in to the network. This makes multi-layer networks powerful in representing nonlinear functions.

Learning rate and momentum: the learning rate and the momentum are the constants chosen to help move the network weights forwards a global minimum for sum squared error (SSE). The learning rate is a constant chosen to help us move the network weights toward a global minimum for SSE. However, what value should it take? How large should the weight adjustments be? In adjusting the learning rates, when it is very small, the weight adjustment tends to be very small. The network probably will take an unacceptably long time to converge. If the learning rate is large, then it will tend the network algorithm overshoot the optimal solution (27).

With respect to the momentum the back-propagation algorithm is made more powerful through the use of this term. Essentially, the momentum term represents inertia. Large values of the

momentum will influence the adjustment in the current weight. A momentum component will help to dampen the oscillations around optimally, by encouraging the adjustments to stay in the same direction (27).

Artificial neural networks have memory that corresponds to the weights in the neurons. Neural networks can be trained offline and then transformed into a process where adaptive learning takes place.

In the analysis of medical data, ANNs have become an alternative to classic statistical methods in recent years. Several ANN techniques were applied in medicine, including methods for diagnosis and prognosis tasks, especially for survival analysis. Most applications of ANNs in medicine refer to classification tasks (39).

2.6 Challenges of Data Mining

Data mining is a relatively new field and there are many challenges to be faced. Extracting useful information from data can be a complicated and sometimes a difficult process. A system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and perform significant worse when a little noise is added to the training set (2).

2.7 Child Health

On this research child health and its focuses, global and national review on child illnesses and its consequence, period of hospitalization, risk factors and impact of child illness on general health preventive and support mechanisms of childhood illnesses are discussed.

Health care professionals and policy makers have focused their attention to measuring and assessing the quality of medical care and methods to improve the quality and the efficiency of health care services. Based on the evaluation made, there is a dramatic decline in under-five illnesses and death in Ethiopia. For example, there was 210 and 104 under- five deaths per 1000 child in 1990 and in 2009 respectively. However in comparison to other developed nations under- five illness and deaths in Ethiopia are among the highest in the world (5).

Hospital admissions in under-five children are mainly from infectious childhood illness that remains an important cause for health service use and health expenditure on these conditions.

Admission is entirely expensive, strains resources, however, can be prevented by giving timely care in the community and/or even at home care, which by contrast is cheaper, avoids the hazards of hospital acquired infection and provides better continuity of care (13).

Hospital admissions vary in different regions of the world. A study conducted in Auckland by using national morbidity database of the New Zealand Health Information Service revealed that, pacific children had the highest hospital admission rates for acute respiratory infections, pneumonia and asthma which are preventable (16). Another study done in England by using The Hospital Episodes Statistics (HES) database also shown unplanned hospital admission rates in young children rose steadily by 22% . The increase in admission rates was greater for common non-infectious cause than infectious cause of admission (13).

Survey conducted in Italy also shown that 30% of the hospital admissions judged inappropriate. Inappropriate admission was seen on variables such as age, sex, type of admission, hour of admission, and location within the stay (17).

Report from Grenn C. Olsen primary General Hospital (NGO Hospital in Ethiopia) in 2005 shown that top five reasons for admission to pediatric ward were pneumonia, severe protein energy malnutrition, kwashiorkor, acute febrile illness, and intestinal obstruction (18). To the contrary preventable diseases are not imminent cause of child morbidity and mortality in economically advanced countries. Thus, morbidity and mortality from infectious childhood illness now dramatically declined and the likelihood of serious bacterial infection in children is now very low (16).

Preventable diseases are not only the cause of illness and admissions but also the major cause of death in under-five children. Pneumonia kills more children than any other disease, yet in developing countries, the proportion of children under five with suspected pneumonia who are taken to appropriate health care providers remains low (6). One survey study done in Gilgel Gibe Field Research Center revealed that neonatal and infant deaths are 38 and 76.4/1000 live births respectively. As indicated in that research, the two most common causes of death during neonatal period were prematurity (24.4%) and pneumonia (22.6%) whereas the top cause of death in post neonatal period were pneumonia (42%), malaria (37%) and acute diarrheal disease (30%) (19).

Potentially preventable diseases are the major cause of illness in Ethiopia. A community based study in urban south western Ethiopia in 2003 revealed that the incidence at diarrheal diseases

was 5.48 episodes per child per year (31). Pneumonia was the most common immediate cause of illness and death (29.7%) followed by acute diarrhea and malaria each contributing for (23%) illnesses and deaths (32). Malnutrition is another common cause of illness in under-five children. According to WHO's new child growth standard released in 2006 underweight and stunting is 34.6 and 50.7 in 2005 respectively.

Another study done in west Gojam shows that 14.8%, 43.2% and 49.2% under-five children are wasted, stunted and underweight respectively (33). However the current trend in under-five child health revealed that there is decrease in morbidity and mortality. According to World Bank report of the 2011, child death is about108 and 69 per1000 in under-five child and infants respectively as compared to 210 and 124 in 1990 (5, 34).

Child illness has a great impact on health status, economic and social support. Study conducted on impact of chronic childhood illness on family stress revealed that parents of children with life threaten and non-life threatening illnesses reported significant differences in social support, perception of the impact of the illness, and coping behaviors (36). Parenting a chronically ill child is a challenge. Having a child with a chronic illness is stressful for any family. Parents of a chronically ill child are often faced with difficulties and decisions that other parents will never have to face (37). Children with a long-lasting physical illness are twice as likely to suffer from emotional problems or disturbed behavior (38).

The factors which influence health are multiple and interactive. It includes the range of personal, social, economic and environmental factors which determine the health status of individuals or populations. Child illnesses may be related with full range of potentially modifiable determinants of health that are primarily related with parents and/or the children themselves such as income and social status, education, employment and working conditions, access to appropriate health services, and the physical environments. These, in combination, create different living conditions which impact on health. According to the 2000 and 2005 Demographic and Health Surveys (EDHS) data marital status, birth order, type of birth, education, household size and sex of household and preceding birth interval are the important proximate determinants for both infant and child morbidity and mortality (35).

2.8 Applications of Data Mining

Nowadays, data are being collected and accumulated at an exponential rate across a wide variety of fields. In response to this huge data, the need for a new tools and techniques to assist humans in extracting novel and useful information and/or knowledge from the large volumes of digital data is mandatory (24).

Data mining is increasingly popular because of the substantial contribution it can make to industries. The telecommunications and credit card companies are one of the intensive users of data mining to detect fraudulent use of their services. Insurance companies and stock exchange are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area in which they are using data mining to predict the effectiveness of surgical procedures, medical tests or medications and identify patterns from patient history (21).

In health industry, with wide spread use of medical information systems including databases, there is an explosive growth in their sizes. The way to extract knowledge in a comprehensible form from the huge amount of data is the primary concern.

Health care analysts and policy makers can learn lessons from the use of KDD in other industries and apply KDD to problems of health care industry (hospitals, insurance companies, physicians and pharmaceuticals companies). Many health service providers are migrating toward the use of computer based patient records and store a large quantity of patient data on test results, medications, prior diagnosis, and medical history. This is a valuable source of information that could be better used by employing data mining techniques.

There are many evidences which approved the application of data mining in different health related problems. Some of these applications are shortly and briefly reviewed as follows:

The application of data mining techniques to a large asthma medical dataset by using CRISP-DM and techniques such as Kohonen’s solf organizing map (SOM) for segmentation, Apriori for association and decision tree for classification revealed that data mining techniques are appropriately applied in mining medical data. The exploration of CRISP methodology suggests that a `one methodology fits all approach is not appropriate, but rather combines to create a hybrid holistic approach (29).

Data mining is also often used in detecting health care fraud. IBM fraud and abuse management system is used for detecting health care fraud and abuse, which ranks as one of the nation's leading law enforcement frustrations (28).

There has also been an explosive growth in biomedical research such as DNA analysis. Data mining finding on DNA analysis has led to the discovery of genetic causes of many diseases and disabilities as well as the discovery of new medicines and approaches for diseases diagnosis, prevention and treatment (2).

CHAPTER THREE

RESEARCH METHODOLOGY

Methodology is the backbone for the research to be carried out. Before one attempts to extract useful knowledge from data, it is important to understand the overall approach. Simply knowing many algorithms used for data analysis is not sufficient for a successful data mining research. The challenge for modern data miners is to come up with widely accepted standards that will stimulate major industry growth. To this end, having predefined methodology in data mining researches is mandatory (15). In this chapter the data mining models, methods, tools and procedures that were used in this research are discussed. Finally an ethical clearance issue was addressed.

3.1 The CRISP-DM Process Model

The model helps to describe procedures that are performed in each step. According to literature in the field of data mining the most popularized knowledge discovery process models are the CRISP-DM and Fayyad et al (15). The main difference between the CRISP-DM and Fayyad et al, KDD model lie in the number and scope of their specific steps. Common features of all models are the definition of inputs and outputs. In Fayyad et al. the number of steps to be followed is nine. It provides detailed technical description with respect to data analysis, but lacks business aspects. In CRISP-DM the numbers of steps followed are six and have good documentation, divide all steps into sub steps. This helps to easily identify all necessary details in the knowledge discovery process.

CRISP-DM is a standard model. It borrowed ideas from the most important pre-2000 models and is the groundwork for many later proposals. It has six steps business understanding, data understanding, data preparation, modeling, evaluation and deployment. That gives recommendations on how to do each task. Furthermore, it is good in documentation and divides all steps in to sub steps (See table 3.1) that make it easier to identify details in the knowledge discovery process (20). Therefore, CRISP-DM model was selected for this research in order to discover knowledge.

Table: 3. 1. The CRISP-DM phases and tasks (20).

Business understanding	**Data understanding**	**Data preparation**	**Modeling**	**Evaluation**	**Deployment**
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next step	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
			Data format		

The details of all steps followed in this research based on the selected model are explained as follows:

3.2 Business Understanding

The effectiveness and efficiency of research work is determined by good understanding of domain/business area. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into data mining problem definition, and designs a preliminary plan to achieve the objectives.

3.3 Data Understanding/Collection

This step was started with initial data collection and familiarization with the data i.e. overall attributes recorded, the values each attribute take, the number of fields and records etc. Specific aims include identification of data quality problems, (completeness, noise, irrelevant records, redundant attributes, missing values, consistency of the records and outliers) initial insights into the data, and detection of interesting data subsets.

3.4 Data Preparation and Preprocessing

All raw data sets initially prepared for data mining are often large; and may be subject to human error. It is expected to have missing values, distortions, misreporting, inadequate sample, etc in the initial data set. To this end preparation and preprocessing of dataset are critical for data mining researchers (1).

Data preparation is typically the least formalized, the most domain- dependent, and the most time consuming part of the knowledge discovery process (29). This step took considerable time of the researcher. In this step the researcher covers all activities needed to construct the final dataset, which constitutes the data that is fed into the data mining tool in the next step. The following sub steps were done in this study.

3.4.1 Data Cleaning

Even though, there are large amounts of data, the amount of completely and correctly filled data may be relatively small. Some of the data mining methods accept missing values and statistically process data to reach a final conclusion (1). In this research the cleaning process was done manually.

There are different alternatives in handling missing values and incompletes. One of the most commonly used method for filling the attributes quickly without too much computation is by replacing all the missing values with the arithmetic mean or the mode with respect to that attribute (2). By adopting the methods specified above, the data cleaning tasks conducted are as follows:

- Missing value for continues/numeric values like age are replaced by the mean value of the field. For nominal variable, the modal (most frequent) value were used
- In some cases, the researcher with experts in the area manually examines samples that have no values and enter a reasonable, probable, or expected value.
- Outliers and noisy field values were handled. For records which are observed having very great difference from the range of values for the specified attributes, a correction was made. This might happen due to the error made when entering the data into the excel format by the users. At this time correction was made by replacing the values with logical estimates. For example, weight of 3 month child is recorded as 65kg. This is a typical outlier. If it is replaced by 6.5kg, it is considered as a logical estimate.

3.4.2 Data Integration and Transformation

Many attributes/fields in the dataset may take more than on value. This might create a great impact on the performance of the algorithm selected. To avoid such problems the researcher has done transformation of attribute values when necessary.

Discretization and concept hierarchy are a necessary preprocessing steps, not just a tool for reducing the data. For example for the decision tree algorithm, reducing the number of values of attributes increases the performance, because, if the number of values of the attribute/features are huge, model building for such data can be difficult and/or highly inefficient. Discretization helps to minimize such problems (15).

A concept hierarchy is defined for the categorical data. Categorical attributes have a finite number of discrete values, with no ordering among the values. The formation of new concept is essential and there are several methods for the generation of concept hierarchies for this type of categorical data. One such method is defining the concept hierarchy by user or expert by specifying a partial or total ordering of the attributes at the schema level. To normalize attributes value the min-max normalization is used. Min-max normalization performs a linear transformation on the original data and it is used because it preserves the relationships among the original data values (2).

3.4.3 Data Reduction and Feature Selection

Irrelevant attributes has a damaging effect on machine learning schemes. Learning with attribute selection, helps to eliminate these irrelevant and consider only most relevant attribute. Some learning methods themselves try to select attributes (decision tree) appropriately and ignore irrelevant or redundant ones, but in practice their performance can frequently be improved by pre-selection (10). Accordingly, selection of relevant attributes was done with experts.

In many practical situations there are far too many attributes for learning. Thus, there should be a mechanism of selecting relevant attributes. The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean. Selection of relevant attributes is important and crucial in data mining. In this research the reduction and selection of attribute was done manually by considering the significance and importance of attributes to the objectives of the research work (10).

Once the relevant attributes are selected, further selection and comparison was made by using information gain on WEKA data mining tools. Dimension reduction helps to balance the training dataset so that the algorithm gets a chance to learn from every group of records.

3.5 Modeling

One of the basic activities in data mining is model building. Classification model was done by using decision tree and ANN algorithm. Decision tree is used to predict categorical variables. Moreover decision tree algorithms are used because they are easy to understand, they are easily converted to a set of production rules, they can classify both categorical and numerical data, but the output attribute must be categorical and there are no a priori assumptions about the nature of the data (22). Decision tree make few passes through to the data. In most cases there is only one pass for each level. In addition decision tree are effective for classification when there is a class label or predictor variables. Therefore, from the variable decision tree algorithm that is supported by the WEKA tool J48, REP tree and Random Tree was analyzed and compared (1).

Artificial neural network (ANN) is another classification algorithm to build a model. It can be used as diagnosis tool in health organizations. Multilayer perceptron algorithm for ANN that is supported by WEKA tool was applied and the result was compared with that of decision tree outputs. In the analysis of medical data, several ANN techniques were applied in medicine, including methods for diagnosis and prognosis tasks, especially for survival analysis. Most applications of ANNs in medicine refer to classification tasks (39).

For learning models, if one of the target variable classes has much lower relative frequency than the other, balancing is recommended. Suppose, a fraud classification model has 100,000 transactions, only 1000 of which are fraudulent, in this example the classification model could simply predict no fraudulent for all transaction and achieve 99% accuracy. Therefore the model seems useless (15). Such a problem is referred as the naïve prediction rate (or naïve error rate). The amount of other data simply swamps the relatively low level of information that is present. In order to get the needed information exposed to the total, the dataset has to be adjusted by manually or automatically by applying SMOTE technique (10).

The choice of a data mining tool is not an easy task. The best tool suit for someone may not be the most advanced tool, or the one that gives the greatest accuracy in prediction. More important than all of these things is identifying the tool that is easy to use, provides acceptable accuracy

(even though not the highest accuracy available), and able to perform all the common tasks in data mining (12).

WEKA, which stands for Waikato environment for Knowledge Analysis, was developed at the University of Waikato in New Zealand. The system is written in Java and distributed under the term of the GNU General Public License (10).

WEKA can handle all the standard data mining problems like regression, classification, clustering, association rule mining and attribute selection. Moreover it is an open source data mining tool. It accepts CVS file formats, and arff. This tool is available for free and can easily be downloaded from site http://www.cs.waikato.ac.nz/-mi/weka/

Raw data can be stored in several formats, such as text, excel or other database types. Converting the data into a format that is understandable by the selected tool should be performed for data preprocessing. In the WEKA tool for example, the data should be stored in the Attribute-Relation File Format (.arff format). First, the data was prepared in excel format as a dataset and saved as comma separated value (CSV) format. WEKA has CSV file loader and can convert it into .arrf format automatically.

3.6 Training

The performance of different algorithms can be measured after a serious of trainings. Data mining tool WEKA has the following options: training set, supplied test set, cross-validation and percentage split training options. From these options 10-fold cross-validation test option was applied for classification task.

10-fold cross validation is good training option because it does not require more data compared to the traditional single percentage split (2/3 training, 1/3 testing) experimentation. The main advantage of 10-fold (or any number of folds) cross validation is to reduce the bias associated with the random sample of the training and holdout data samples by repeating the experiment 10 time, each time using a separate portion of the data as holdout sample (30).

3.7 Analysis and Evaluation

In this step, the researcher analyzed the result of the algorithms by a selected tool. The results of the algorithms were also compared with the ideas of experts and users.

The analysis and evaluation was conducted by observing the confusion matrix. The confusion matrix is simply a square matrix that shows the various classification and misclassifications of the model in a compact area. The columns of the matrix correspond to the number of instances

classified as a particular value and the rows correspond to the number of instances with actual classification. Exploratory data analysis technique was applied to visualize the result. This technique is used to present results in graphs and tabular formats. For modal evaluation and selection, even if the above parameters can be used, the suggestion, choice and opinion of experts and users were taken into consideration.

Confluence of results is necessary in modal selection. The main idea in modal selection is that the researches and analyst should not depend on one kind of methods (27). Based on that, the models are selected by comparing different algorithms performance in terms of accuracy, true positive rate, false positive rate, the ROC (receiver operating characteristics), users comment and suggestion.

3.8 Deployment

At this step the researcher plans for the dissemination of the research outputs. Primarily, producing final report in order to, communicate the outputs with the users at domain area and similar settings. This model also enhances decision making at different level of health organizations and NGO's as well. Therefore, publication of findings in health journals and presenting findings in health conferences is planned.

3.9 Ethical Clearance

To fulfill the ethical issues the researcher has secured ethical clearance letter from the Addis Ababa University and other respective organizational managers. Oral or written consent was taken for any cases that needed confidentiality.

CHAPTER ONE

EXPERIMENTATION and DISCUSSION of RESULT

Data mining algorithms are found to be useful for automatic classification of data. While performing these tasks, some algorithms might work better than others when running one type of data as compare to the others. Thus, finding the best type of algorithm that best suit the specific type of data for better result is an interesting and time consuming task. For this study decision tree and ANN algorithms are analyzed. As it is indicated in the methodology part of this study the tool that is used for the classification is WEKA.

4.1 Business Understanding

Hadiya zone is one of the largest Zones in south nation nationalities peoples region (SNNPR), Ethiopia. It is located 232kms south west of Addis Ababa. According to the national census of 2007 the total population of the Zone is estimated to be around 1,371,625 from these infants and under- five children comprises of 48,007 and 213'971 respectively. Hossana hospital recently called NEMM Hospital, is only one Public Hospital, under SNNPR and federal government, and is governed by Zonal administration which is located in Hossana town. It serves the zonal population and also people nearby the zone as well.

The Hospital has four major inpatient departments (medical, surgical, pediatric, Gynecology/obstetrics), outpatient department and operation room. Under- five children are treated/take services in under-five OPD and pediatric ward. Children older than five-years and surgical cases are managed in other OPDs and surgical ward according to the cases and the hospital's management protocol. According to the report of the hospital, the major causes of under-five OPD visits are pneumonia, malaria, diarrhea, malnutrition, and other febrile illnesses.

The information of children in under-five OPD visit is recorded in daily bases in integrated under-five registration log book. Some of the records contained are age, sex, address, HMIS disease classification, HIV test, outcome of admission, danger signs, other unspecified signs and symptoms, presenting complaints (fever, cough, diarrhea, etc), immunization status, vitamin A supplementation , weight, weight for age, and length of stay. Based on their working procedure outcome of under-five visit (admitted, not admitted) can be taken as a class label and other attributes are checked whether they can yield better result in data mining. There are 15,824 records of children in under- five OPD since June 7, 2006.

4.2 Data Understanding

4.2.1 Data Selection Process

The dataset for this study was obtained from Hossana Hospital recently named as Nigist Eleni Mohammed Memorial Hospital. It was stored in integrated under-five registration log book in under-five OPD. The main reason of storing the data was entirely for case report.

The data used for this study was the records kept from June 7, 2007 to February 7, 2012. Within this period, there was a total of 15,824 under-five OPD visits ranging from 5- 19 cases and on average there was 13 cases visiting under-five OPD per day. Therefore the data collection for this research begun by coding these data into MS excel format.

4.2.2 Basic Data Distribution

The dataset were organized in columns and rows where the columns represent an attribute and the rows represent single records of under-five OPD visits cases. In general there was 64 attributes (columns) and 15,824 records (rows) in under-five OPD. Out of these under-five visits 1,264 records are stored in integrated under-five registration log book of birth to 2 months and the rest 14,560 records are stored in integrated under-five registration log book of 2 months to 5 year.

Regarding to selection of best attributes, some attributes recorded in integrated under-five registration log book of birth to 2 month and 2 month to 5yrs are different. Thus, by taking the opinion of experts those attributes that are similar and mostly supportive for the diagnosis of the diseases are selected. Therefore, final data preprocessing was done on 28 attributes and 15,824 under-five OPD visits.

The outcome of under-five OPD visit is either of getting treatment at OPD (not admitted) or to be treated in pediatric ward (admitted). Accordingly, 28 selected attributes, data types, number of unique value the attributes take, and the number of missing values are showed in table 4.1 below.

4.3 Data Preprocessing

As shown in table 4.1, these selected attributes contained missing values. Other things that are seen in the raw data are incompleteness and noise. There are also fields that are redundant and containing derived values. Therefore, such unnecessary attributes are removed and the remaining attributes are considered for further preprocessing.

Table: 4.1 List of attributes selected for the study from NIgist Eleni Mohammed Memorial Hospital from Jun7, 2007 to Feb7, 2012.

S.n	Attribute	Data Type	Description of attributes	missing
1	Address	Categorical	Geographical distribution of the illnesses	77 0.49%)
2	Age	Numeric	The age of under-five child	4 (0.03%)
3	Sex	Categorical	The sex of under-five child	8 (0.05%)
4	Weight	Numeric	The weight of under-five child	7 (0.05%)
5	Visit	Categorical	Under-five OPD visit status of the under-five child	23(0.1%)
6	Cough	Categorical	complaint of under-five child to seek medical care	14(0.09%)
7	Fever	Categorical	complaint of under-five child to seek medical care	2(0.13%)
8	Diarrhea	Numeric	complaint of under-five child to seek medical care	29(0.18%)
9	Vomiting	Numeric	complaint of under-five child to seek medical care	17(0.11%)
10	Abd/ pain	Numeric	complaint of under-five child to seek medical care	6(0.04%)
11	TIPA	Numeric	complaint of under-five child to seek medical care	18(0.11%)
12	Other	Numeric	complaint of under-five child to seek medical care	32(0.2%)
13	Una/feed	Categorical	complaint of under-five child to seek medical care	13(0.08%)
14	Convul.	Categorical	complaint of under-five child to seek medical care	7(0.05%)
15	Let/uncon.	Categorical	complaint of under-five child to seek medical care	15(0.1%)
16	F/breath	Categorical	complaint of under-five child to seek medical care	11(0.07%)
17	Stridor	Categorical	complaint of under-five child to seek medical care	3(0.02%)
18	C/drawing	Categorical	complaint of under-five child to seek medical care	5(0.03%)
19	Bul/font.	Categorical	complaint of under-five child to seek medical care	6(0.04%)
20	R/irritable	Categorical	complaint of under-five child to seek medical care	1(0.01%)
21	B/dinstool	Categorical	complaint of under-five child to seek medical care	2(0.13%)
22	Sun/ eyes	Categorical	complaint of under-five child to seek medical care	11(0.07%)
23	Skin pinch	Categorical	complaint of under-five child to seek medical care	5(0.03%)
24	W/ for age	Categorical	complaint of under-five child to seek medical care	3(0.02%)
25	Oral trash	Categorical	complaint of under-five child to seek medical care	2(0.13%)
26	I/status	Categorical	Immunization status of under-five child	34(0.22%)
27	OPD visit	Categorical	Determining OPD visit as admitted or not admitted	0(0%)
28	HMIS	Categorical	Diagnosis of under-five child made at OPD	0(0%)

As mentioned in the methodology part of this study, the dataset that are selected and prepared for the mining purpose should have initial feature (relevant attributes). The selection of relevant attribute was done by working with experts. Therefore, the next part of preprocessing is handling missing values, incorrect values, noises and other irrelevant values for the attribute/fields selected.

Missing values are observed in all selected attributes except the outcome of under-five visit and HMIS disease classification. Usually missing values are replaced either by mean value or modal value for numeric and nominal attributes respectively. For numerical variables like age and weight the missing value is replaced by the mean value of age and weight respectively. Interestingly the tool WEKA replaces missing value by mean value. For all nominal variables like sex, cough, immunization status, oral trash, etc the modal value is filled manually.

When it comes to outliers and noise field values, correction is done manually. Data is entered incorrectly for some attributes which created noise and contained outlier. For some other attributes whose value is unique categorical, but incorrectly values other than the mentioned unique values are obtained. For example the sex of the child may be male (M) or female (F) and that is unique categorical value. But some of it is recorded as mm, ff, fm, & mf and these all create noise. The weight of the child is recorded as 85kg. This is a typical outlier and should be corrected. Such problems are handled and correct values are substituted based on the methods described in the previous chapter. The detail of how the attributes were handled in removing outliers and noise field values was described in the following section.

Age: - An attribute age is recorded in days, weeks, months and years. This makes it difficult for analysis. Based on the opinion of the experts, some disease classifications and management of child illness relies on the age classification of the children. For example neonatal sepsis, neonatal tetanus etc are illnesses that occur at neonatal period. Child immunization takes more attention at infantile period. Integrated management of neonatal and childhood illnesses (IMNCI) management in under-five OPD also relies on managing children less than two months in one category and children older than 2months to five years in another category. In addition to that this transformation helps to identify which developmental age group is more times visiting OPD for which illness and also it takes the attention of care providers in managing illness related with specific age category. Therefore, taking the name of child's development periods is relatively

appropriate in classifying the age of children. Table 4.2 displays the age transformed for children.

Table: 4.2 Age transformed for children based on the child's developmental period.

S. No	Child's age	Child's developmental period
1	0-28days	Neonate
2	29days-1year	Infant
3	1year to five years	Child

Address: - Address reflects the geographical distribution of the children visiting the under-five OPD and those children admitted and discharged from pediatric ward. As the Hospital is located in Hadiya Zone, the people visiting the Hospital were primarily from woredas of Hadiya Zone and other neighbor zones. Accordingly, those children visiting from Hadiya zone are recorded as their respective woredas such as Hossana, Lemo, Analemo, Misha, Gombora, Soro, Duna, Shashogo,and Gibe. Meanwhile those children visiting from Lera, Silte, Worabe, Wulbareg, Sankura, etc are recorded as Silte Zone, those from Azernet, Endagagn, Wolkite, etc are recorded as Gurage Zone and those from Angecha, Hadaro, Doyogena etc are recorded as KT Zone.

Sex: - The sex of children visiting under-five OPD and pediatric ward is also an important attribute selected for this study. It is recorded as either "m" or "f". The letter "m" which is transformed to male and female, but missing values is replaced by WEKA tool for its modal value.

Visit: - Another attribute selected is the frequency of visit of the children to under-five OPD or pediatric ward. Those children visited for the first time are recor ded as "initial" and children visiting more than one times are recorded as "follow up" visit, but missing values are replaced by WEKA tool for its modal value.

Weight: - Weight is also an important attribute selected for this study, because it is direct reflection of the child's health and nutritional status. In relation to age, it is a normal child growth indicator because as the age of child increases the weight also increases at a proportional rate termed as appropriate weight for specified age. To the contrary, the child's weight which is not proportional to the age of child is termed as large weight for age, low weight for age or very low weight for age. As it is continuous numeric variable, it automatically descritized into five

bins by using WEKA tool for analysis and respective model building. The summary of some selected attributes is shown on table 4.3.

Presenting Complaints of the Child: - The next attribute selected for this study was the presenting complaint of the child (symptoms) and physical findings observed by health care providers (signs). Under this attribute the symptoms presented on the child (subjective complaint of the patients) that urge the patient to visit under-five OPD were recorded. Group of signs and symptoms presented on children are important clinical findings to know the specific cause of illnesses that resulted in under-five OPD visit. So that, which clinical findings are related with the specific diagnosis is an important area of this study. Therefore, in the following section clinical signs and symptoms, those are selected as best attribute for this study were briefly described.

Before the selection of best signs and symptoms, knowing what clinical signs and symptoms are recorded in two integrated under-five registration log books (birth to 2 month and 2 month to 5yrs) in under-five OPD is a crucial issue. Clinical signs and symptoms such as cough, fever, diarrhea, vomiting, abdominal pain/cramp, trauma/injury/accident/poisoning, unable to drink/breast feed, convulsions, lethargic/ unconscious, fast breathing, stridor, bulged fontanelle, irritable/restless, sunken eyes, oral trush, blood in stool, low weight for age , chest indrawing are recorded in both birth to 2 month and 2 month to 5yrs integrated under-five registration log books. Clinical signs and symptoms such as nasal flaring, grunting, gestational age, movement less than normal, jaundice, breast feeding, supplementary feeding are recorded in birth to 2 month integrated under-five registration log book only whereas stiff neck, red eye, mouth ulcer, ear pain/discharge, foot edema, MUAC <10cm, pallor, appetite are recorded in 2 month to five year integrated under-five registration log book only.

For the selection of attributes, the opinion of the experts was taken as an important input. Children may complain for cough, fever, diarrhea, vomiting, abdominal pain/cramp, trauma/injury/accident/poisoning, unable to drink/breast feed, convulsions, lethargic/unconscious, fast breathing, stridor, bulged fontanelle, irritable/restless, sunken eyes, oral trush, blood in stool, low weight for age , chest indrawing and others. All these, clinical signs and symptoms are grouped into subjective complaints or objective findings that may results in under-five OPD visit and/or pediatric ward admission.

Table: 4.3. Summary of some selected attributes, their old records and the new value records it takes for predicting cause of under-five children admission

Attributes	**Old values recorded**		**New value it takes**	**Remark**
Age	Age in days, weeks, months, and years		Neonate, infant, child	Nominal
Sex	m ,f		Male, Female	Nominal
Visit	initial, follow up		Initial, Followup	Nominal
Address	Hadiya zone	hossana, lemu, analemu, duna, gombora, misha , gibe, soro , shashogo, badewacho	Hossana, Lemu, Analemu, Duna, Gombora, Misha, Gibe, Soro, Shashogo, Badewacho	Nominal
	Neighbor Zones	lera, silte, worabe, wulbareg, sankura	Siltezone	
		doyogena, angecha, hadaro, obichaka	KTzone	
		azernet, endagagn,wolkite	Guragezone	
weight		Weight of children in kg	Weight of children in kg	Numeric continuous

By interpreting all these findings and professionals judgment HMIS disease classifications was termed for each under-five child visit or admission. Hence, all mentioned clinical findings that are recorded in both log books were taken as a best attribute for this study as depicted in table 4.4.

Table: 4.4 Details of selected clinical signs and symptoms as an attribute to predict the outcomes of under-five OPD visit.

S.n	Attributes	Description	Value it takes	Remark
1	Cough	Presenting complaint of the child	Yes or no	2 categorical
2	Fever	Presenting complaint of the child	Yes or no	2 categorical
3	Diarrhea	Presenting complaint of the child	Yes or no	2 categorical
4	Vomiting	Presenting complaint of the child	Yes or no	2 categorical
5	Abd pain/cramp	Presenting complaint of the child	Yes or no	2 categorical
6	Trauma/injury/accident/poisoning	Presenting complaint of the child	Yes or no	2 categorical
7	Unable to drink/breast feed	Presenting complaint of the child considered as danger sign	Yes or no	2 categorical
8	Convulsion history/now	Presenting complaint of the child considered as danger sign	Yes or no	2 categorical
9	Lethargic/unconscious	Presenting complaint of the child considered as danger sign	Yes or no	2 categorical
10	Fast breathing	Presenting complaint of the child	Yes or no	2 categorical
11	Stridor	Presenting complaint of the child	Yes or no	2 categorical
12	Chest indrawing	Presenting complaint of the child	Yes or no	2 categorical
13	Bulge fontanelle	Presenting complaint of the child	Yes or no	2 categorical
14	Restless/irritable	Presenting complaint of the child	Yes or no	2 categorical
15	Blood in stool	Presenting complaint of the child	Yes or no	2 categorical
16	Sunken eyes	Presenting complaint of the child	Yes or no	2 categorical
17	Skin pinch	Presenting complaint of the child	Normal, low, veryslow	3 categorical
18	Weight for age	Presenting complaint of the child	AWFA,VLWA, LWFA	3 categorical
19	Oral trash	Presenting complaint of the child	Yes or no	2 categorical
20	others	Complaint of the child not specified in the above complaints	Yes, no	2 categorical

Immunization Status: - Immunization status of the child is another important risk factor in assessing the health status of the children. In Ethiopian setup, currently children are immunized against eight targeted diseases (tuberculosis, polio, diphtheria, tetanus, whooping cough, hepatitis B virus, hemophylus influenza type B, measles) in all governmental health organizations and some NGOs. In addition to that, very recently immunization against streptococcus pneumonia is introduced. Therefore, children immunized against targeted diseases are assumed to be protected from those diseases. In this sense knowing the immune status of the child in relation to child's disease condition is one of the crucial issues. Accordingly, immunization status of the child is selected as attribute for this study.

Immunization status is recorded as completed, up to date, defaulted or not started. Immunization status is completed to mean the child is completely vaccinated against targeted diseases, up to date means the child has started vaccination but not completed, defaulted means the child has started vaccination but not continued the next schedules to complete it and not started means the child yet not started immunization. A detail of the immunization status is shown in table 4.5.

Table: 4.5 Details of immunization status.

Attributes	**Description**	**Value it takes**	**Remark**
Immunization status	The child's immune status against eight targeted diseases	Completed, uptodate, defaulted, notstarted	4 categorical

HMIS Disease Classification: - Under-five children visit under-five OPD for their illnesses. Ministry of health classifies these illnesses as HMIS diseases. HMIS diseases are termed as HMIS admission disease classification based on the child's admission diagnosis made in under-five OPD. It describes for what illness the child was admitted to the pediatric ward. For example, HMIS admission disease classification is malaria means the child is admitted to pediatric ward for the illness malaria and takes medical care in the ward. Therefore, HMIS admission disease classification is the diagnosis made in under-five OPD for the illness that child to visit under-five OPD. Hence, it is one of the attributes selected for this study.

A total of 116 HMIS admission disease classifications were recorded with the highest frequency of pneumonia (4064), 25.68% and the other 2 diagnosis each consists of (3) 0.01%. This shows

that, there was high difference between common cases and rare illnesses. From these points of view it needs some adjustments and the experts' involvement was needed. So that, by giving primary consideration for the objective of the study and the impact of using 116 values for one attribute on algorithms to be used, the following work was done by involving experts.

Accordingly, representing some of the values by some other higher concepts was done by considering the relative prevalence of the illnesses, illnesses affecting the organ system of the body, those illnesses that affect the same organ but having different diagnosis, diagnosis that were given different names based on its severity and illnesses of very rare diagnosis. Hence, a total of 53 categories were selected for the attribute HMIS admission disease classification.

The details of HMIS admission disease classification selected as an attribute and value it takes for predicting cause of under-five children admission are shown in table 4.6.

Outcome of Under-five OPD Visit: - Dependent attribute (outcome variable) for this study is the outcome of under-five OPD visit. Based on the findings assessed at under-five OPD, the Doctors or other health professionals reached at the diagnosis of the child's illness. These illnesses are assigned into HMIS disease classification, where it needs immediate treatment under direct observation of physician and nursing care (admitted) or provided appropriate treatment at OPD level or else referred to other health facility for further management(not admitted). Therefore, prediction of cause of under-five children admission is possible at this point by taking different attributes as risk factors and taking admitted or not admitted as an outcome variable. Hence, early prediction of admissions resolve related problems such as resource management as discussed on the first chapter in this study.

4.4 Data Reduction and Feature Selection

Inputs from domain expert are essential in data reduction and feature selection process. Because expertise in assessing child health condition and management is not trivial, their involvement in the preprocessing step is more essential. Thus, candidate variables were selected primarily by taking the research objectives and bases of experts influence in assessing child health condition and management. The reduction process applied on attributes is described as follows.

1. Attributes like serial number, medical registration number, date of visit, date of admission, data of discharge, amount of charge, amount paid, voucher, counsel

Table: 4.6. HMIS disease classification

S. N	Attribute		
	Old value	New value	Value it takes
1	Malaria	malaria	Malaria
2	TAPF	Trauma, accident, poisoning,fructure	TAPF
3	Helminthiasis	helminthiasis	Helminthiasis
4	Burn	burn	Burn
5	Severe pneumonia	severepneumonia	Severepneumonia
6	UTI	UTI	UTI
7	Skin infection	skininfection	skininfection
8	Meningitis	meningitis	Meningitis
9	Tonsillitis	tonsilitis	tonsilitis
10	Diarrhea	diarrhea	diarrhea
11	Foreign body	foreignbody	foreignbody
12	Constipation	constipation	constipation
13	Severe acute malnutrition	SAM	SAM
14	Typhoid fever	Typhoid fever	Typhoidfever
15	Nephritis	Nephritis	Nephritis
16	Conjunctivitis	Conjunctivitis	Conjunctivitis
17	Hernia	Hernia	Hernia
18	Genitalia problems	Genitalia problems	Genitalia problems
19	Infected wound	Infected wound	Infectedpneumonia
20	Cellulitis	Cellulitis	Cellulitis
21	Liver and GI problems	Liver and GI problems	LGIP
22	Pneumonia	Pneumonia	Pneumonia
23	Dysentery	Dysentery	Dysentery
24	Tumor	Tumor	Tumor

25	Neonatal sepsis	Neonatal sepsis	Neonatalsepsis
26	Lymphadinitis	Lymphadenitis	Lymphadenitis
27	Others	Others	others
28	Intestinal obstruction	Intestinal obstruction	Intestinalobstruction
29	Epilepsy	Epilepsy	Epilepsy
30	Mastoditis	Mastoditis	Mastoditis
31	Rectal prolapsed	Rectal prolapsed	Rectalprolapse
32	Measles	measles	Measles
33	AFI	Acute febrile illness	AFI
34	ORTI	Other respiratory infections	ORTI
35	Seizure	Seizure	Seizure
36	Tuberculosis	Tuberculosis	Tuberculosis
37	Prematurity	Prematurity	Prematurity
38	Anemia	Anemia	Anemia
39	Epistaxis	Epistaxis	Epistaxis
40	Pyomyocitis	Pyomyocitis	Pyomyocitis
41	RIV	RVI	RVI
42	RF	RF	RF
43	Heart problems	Heart problems	Heartproblems
44	Meconium aspiration	Meconium aspiration	Meconiumaspiration
45	DKA	DKA	DKA
46	Tetanus	Tetanus	Tetanus
47	Arthritis	Arthritis	Arthritis
48	Abscess	Abscess	Abscess
49	Sepsis	Sepsis	Sepsis
50	Hypothermia	Hypothermia	Hypothermia
51	Congenital deformity	Congenital deformity	Congenitaldeformity
52	Unlikely bacterial infection	Unlikely bacterial infection	Unlikelybacterialinfection
53	Umbilical bleeding	Umbilical bleeding	Umblicalbleeding

mother, treatment given, defined date to return follow up were removed because they have no value in predicting length of stay. In addition attributes like amount of charge, amount paid, voucher and treatment contain too many missing values (more than 80%).

2. Attributes such as HIV test offered, HIV test performed, temperature in degree centigrade, vomits everything, breathing problem and birth weight, were removed because all these attributes are redundant.
3. Attribute like duration of cough, respiration rate in minutes, risk area of malaria, check Vitamin A supplementation in last six months, check mebendazole supplementation, and other problems were removed because it contains too many missing values (more than 80%).
4. Attributes like drinking poorly, history of measles last six months, generalized skin rash, deep mouth ulcer, cough/runny nose/red eyes, tender swelling behind the ear, visible wasting, MUAC <10cm, edema of both feet, poor appetite, history of persistent diarrhea/pneumonia/dysentery, HIV related disease 1,2,3, bilateral parotid enlargement, enlarged lymph nodes, were recorded in two months to five year integrated registration log book only. Whereas gestational age, nasal flaring, grunting, condition of umbilicus, skin pustules, jaundice, movement less than normal, assessing infant feeding habit such as infant breast feeding, frequency of breast feeding in 24 hours, switching from breast to breast before completing one breast, start other feeding and its frequency, feeding by using, feeding during illness, inappropriate placement of milk and fluid, insufficient replacement of milk, replacement milk unhygienic, mixed feeding , bottle feeding, positioning during breast feeding, attachment during breast feeding and sucking effectively were recorded in birth to two months integrated registration log book only. Thus, all were omitted because of lack of uniformity in records and the values they take.

Finally as described above in the preprocessing part of this study, 28 attributes were selected. Annex A- describes the selected attributes, the values that the attribute takes, as well as number and percent composition of the values.

AS indicated in annex-A among under-five children who visited under-five OPD, nearly 70%of children were at the age category greater than 7.5months, 57.29% were male and 60.94% were visiting from Hossana town followed by Lemu woreda (18.47%). More than 95% of admissions to the pediatric ward were initial visits.

Regarding to the presenting complaint of children, 48.27% of children were presented with cough and 44.72% of children had fast breathing. 30.38% of children visited to under-five OPD were presented with fever, 19.89% with vomiting, 19.31with diarrhea and 14.74% had chest indrawing whereas 17.9% of complaints were other unspecified complaints. 96.85% of children who visited under-five OPD had almost appropriate weight for their age.

Pneumonia (25.68%) and severe pneumonia (15.8%) only accounts for 41.48% of under-five OPD visits followed by diarrhea which comprises of 17.43% of cases. 10.62% of under-five OPD visits were due to tonsillitis, 3.14% were due to other respiratory infections such as croup, asthma, bronchitis whereas neonatal sepsis accounts for 2.94% and 2.92% were due to malaria. Severe acute malnutrition comprises of 2.42% of under-five OPD visits, whereas 2.07%, 1.18% and 1.04% of under-five OPD visits were due to helminthiasis, tuberculosis and meningitis respectively. Only 14.78% of under-five OPD visits were due to other illness other than pneumonia, severe pneumonia, diarrhea, tonsillitis, croup, bronchitis, asthma, severe acute malnutrition, malaria, meningitis, helminthiasis tuberculosis and neonatal sepsis.

What is understood from these points is that there was a great variation of values among attributes and needs balancing of values for each attributes. Interestingly the WEKA tool has Synthetic Minority Over-sampling TEchnique (SMOTE) where it automatically balances the dataset. Therefore, the SMOTE technique was applied to balance the data before the experimentation done.

4.5 Classification sub Phase

This experimentation is the major task to classify a data item into one of the several predefined classes. Then the classification outputs obtained by experimenting on different algorithms were compared. Finally, the best selected model will be used for predicting cause of under-five children admission to pediatric ward.

Decision variables were used for predicting cause of admission of under-five children to pediatric ward. An outcome of under-five OPD visit with admitted and not admitted values is used as a decision variable. The total dataset (28 attributes and 11,774 records) were used to construct the decision tree and artificial neural network (ANN). Different algorithms were implemented by using the WEKA tool, which is evaluated by 10 fold cross-validation test option.

4.5.1 Decision Tree Model Building Experiment

For this classification experiment, 11,774 records with 28 attributes of children visited under-five OPD was implemented. In order to improve the accuracy and the performance of the algorithm, attributes were selected by using their information gain. For example, the WEKA attributes selection information gain algorithm applied on 10 fold cross-validation modes revealed that, the output of the result based on their rank are: HMIS disease classification, chest indrawing, unable to feed or drink, skin pinch, stridor, sunken eyes, weight for age, fast breathing, lethargic or unconscious, convulsion history or convulsing now, weight, vomiting, cough, diarrhea, restless or irritable, age, bulged fontanelle, visit, fever, blood in stool, other unspecified signs and symptoms, trauma/injury/accident, abdominal pain/cramp, and sex for children visited under-five OPD.

Decision tree experimentation was conducted, where three different kinds of algorithms (J48, Random tree and REP tree) were tested in WEKA. For this experimentation, first all 28 attributes of children who visited under-five OPD were tested which is followed by ignoring two attributes (sex and abdominal cramp/pain). Then another two attributes (trauma/injury/accident and other unspecified sign and symptoms) were ignored. Third experimentation was applied by ignoring another two attributes (blood in stool and oral trash). The next experimentation was implemented by ignoring an attributes address and immunization status. Final experimentation was applied by ignoring another four attributes age, weight, cough, and visit. The details of the experimentation outputs were summarized on table 4.7.

As shown in table 4.7, for the decision tree the models are built for each classifier. When models are compared, in terms of accuracy, time required for building the model, size of tree, WTPR, WFPR and WROC, model built by running J48 algorithm has relatively higher accuracy, WTPR, WFPR and WROC than REP tree and Random tree. REP tree has small number of tree sizes, time required to build the model is also shorter than J48. When compared to REP tree and J48 the model built by implementing random tree algorithm is faster but lower accuracy, WTPR, WFPR, WROC and also the size of the tree is not manageable. Therefore, J48 performance is better than REP tree and random tree.

Table: 4.7 Performance of classification algorithms output.

Algorithm	No. of attributes	Tree size	Time elapsed in second	Correctly classified instances	Accuracy in %	WTPR	WFPR	WROC
J48	28	375	0.48	11291	95.90	96.5	3.5	0.998
	26	401	0.45	11262	95.86	96.4	3.7	0.996
	24	386	0.38	11281	95.87	96.3	3.7	0.989
	22	386	0.37	11281	95.87	96.3	3.7	0.989
	20	223	0.34	11230	95.80	96.3	3.7	0.989
	16	153	0.23	11111	94.77	94.7	5.3	0.99
Random tree	28	2160	0.08	11274	95.75	96.2	3.9	0.980
	26	2450	0.05	11291	95.90	96.3	3.8	0.982
	24	2475	0.09	11255	95.59	96.2	3.9	0.981
	22	2915	0.11	11280	95.80	96.2	3.8	0.980
	20	1368	0.13	11238	95.45	96.0	4.0	0.981
	16	909	0.05	11156	94.75	94.8	5.3	0.981
REP tree	28	373	0.27	11291	95.90	96.3	3.7	0.990
	26	354	0.42	11286	95.86	96.3	3.7	0.989
	24	352	0.22	11284	95.84	96.2	3.8	0.987
	22	352	0.22	11284	95.84	96.2	3.8	0.988
	20	178	0.36	11253	95.58	96.1	3.9	0.988
	16	109	0.13	11152	94.72	94.3	5.7	0.983

4.5.2 Artificial Neural Network (ANN) Model Building Experiment

Artificial neural network (ANN) is another classification algorithm to build a model. ANN can be used as diagnosis tool in health organizations. All 11,774 records with 28 attributes of children who visited under-five OPD and 5,845 records with 30 attributes of children admitted to pediatric ward was implemented in this classification experiment. The algorithm supported by WEKA for this experiment is Multilayer perceptron. ANN accepts inputs that are in binary form. The normalization process was handled by using WEKA tool. All attributes except the target were normalized.

Once the data is ready, the experimentation was conducted by using the outcomes of under-five OPD visit (admitted and not admitted) as target attribute. Similar to decision tree, different attribute sets were tested. For example, first all 28 attributes of children visited under-five OPD were tested which is followed by ignoring two attributes (sex and abdominal cramp/pain). Then another two attributes (trauma/injury/accident and other unspecified sign and symptoms) were ignored. Third experimentation was applied by ignoring another two attributes (blood in stool and oral trash). The next experimentation was implemented by ignoring an attributes such as address and immunization status. Final experimentation was applied by ignoring another four attributes such as age, weight, cough, and visit.

Different parameters were also arranged to see the performance of the algorithms. For example, the learning rate and number of hidden layers were modified and the result for which the performance is best was selected. The details of the experimentation outputs were summarized on table 4.8.

The result of multilayer perceptron algorithm for ANN classifier revealed that, when the number of attributes experimented decreases, the time required for building the model and the number of nodes also decreases by assigning learning rate and momentum as constant. The accuracy (the instances of the children that are correctly classified) of the model is also greater than 94.5%. WTPR, WFPR and WROC of the classifier are also higher. Meanwhile when 16 attributes are tested by assigning nodes 8 and learning rate and momentum as 0.3 & 0.2, the time required to build the model is much lower than the previous tested value, but the accuracy is slightly lower than the previous test values. For the same attributes when the node is assigned at 8 and learning rate and momentum at 0.5 & 0.4, the accuracy still decreased slightly but the time required for building the model increased for about 1.94 seconds.

In general the number of nodes and the time required for building the model is very high. Moreover, the outputs of the algorithms are too difficult to interpret for domain experts.

4.5.3 Comparison of Models

As observed in the previous sections, for classification models built, different attributes and parameters were set in order to get the classifier that has good performance. So that the classifier that built the models with high performance will be applied in the actual settings. The other important thing that was considered during comparison of models is the simplicity of the model

Table 4.8 Outputs of multilayer perceptron at learning rate 0.2 and momentum 0.3

No of attributes	**Time in (seconds)**	**No of (nodes) Hidden layers**	**instances classified**	**Accuracy in %**	**WTPR**	**WFPR**	**WROC**
28	2567.88	53	11164	94.82	94.8	5.2	0.977
26	3797.87	51	11201	95.13	95.1	4.9	0.975
24	1418.18	50	11191	95.05	95.0	5.0	0.977
22	1342.15	49	11194	95.07	95.1	5.0	0.976
20	950.23	40	11184	94.99	95.0	5.0	0.980
16	858.6	38	11138	94.60	94.6	5.4	0.979
16	186.2	8	11135	94.57	94.6	5.5	0.981
16	188.14	8	11101	94.28	94.3	5.8	0.976

for users and applicability of the model in actual working areas. Therefore, for the comparison of the models primarily the accuracy, WTPR, WFPR, WROC of the classifier and the time required for building the models are taken. The details of comparison of classification models were summarized on table 4.9.

As presented in table 4.9 for the decision tree and ANN the models are built for each classifier. When these models are compared, in terms of accuracy, WFPR, WFNR, WROC and time required for building the models, the model built by running J48 algorithm has relatively higher accuracy, WTPR, WFPR, and WROC than REP tree, Random tree and multilayer perceptron algorithms.

REP tree has small number of tree size than J48 and Random tree. J48 has also very much small number of tree size than Random tree. Random tree algorithm requires shorter time for building the models than all other algorithms however, the size of the tree is not manageable. In other words Multilayer perceptron requires the longest time period for building the models of all algorithms.

Generally when the performances of all algorithms are compared, J48 performance is better than REP tree and random tree. When J48 algorithm is compared to that of multilayer perceptron, J48 algorithm has relatively higher accuracy, WTPR, WFPR, WROC and performs much faster than multilayer perceptron. Therefore, based on that when the performance of the classifier of the

Table: 4.9 Comparison of classification models

Algorithms	**No. of attributes**	**Time elapsed in second**	**Accuracy in %**	**WTPR**	**WFPR**	**WROC**
J48 Decision tree	28	0.48	95.90	96.5	3.5	0.998
	26	0.45	95.86	96.4	3.7	0.996
	24	0.38	95.87	96.3	3.7	0.989
	22	0.37	95.87	96.3	3.7	0.989
	20	0.34	95.80	96.3	3.7	0.989
	16	0.23	94.77	94.7	5.3	0.99
Random tree decision tree	28	0.08	95.75	96.2	3.9	0.980
	26	0.05	95.90	96.3	3.8	0.982
	24	0.09	95.59	96.2	3.9	0.981
	22	0.11	95.80	96.2	3.8	0.980
	20	0.13	95.45	96.0	4.0	0.981
	16	0.05	94.75	94.8	5.3	0.981
REP tree decision tree	28	0.27	95.90	96.3	3.7	0.990
	26	0.42	95.86	96.3	3.7	0.989
	24	0.22	95.84	96.2	3.8	0.987
	22	0.22	95.84	96.2	3.8	0.988
	20	0.36	95.58	96.1	3.9	0.988
	16	0.13	94.72	94.3	5.7	0.983
Multilayer perceptron ANN	28	2567.88	94.82	94.8	5.2	0.977
	26	3797.87	95.13	95.1	4.9	0.975
	24	1418.18	95.05	95.0	5.0	0.977
	22	1342.15	95.07	95.1	5.0	0.976
	20	950.23	94.99	95.0	5.0	0.980
	16	858.6	94.60	94.6	5.4	0.979

decision tree is compared to that of ANN, decision tree models has higher performance than ANN.

4.6 Decision rules of J48 algorithm and its interpretations

From the outputs of J48, 35 interesting rules are obtained. According to that unable to feed or drink is the first attribute in determining the admission of children to pediatric ward followed by weight of the child. Among all rules obtained, four of the first rules generated and their interpretations are discussed in the following section:

If (unable to feed/drink=yes) then the outcome of under-five OPD visit = admitted

- It is obviously known that in the working procedure at under-five OPD, to admit children primarily the HMIS admission disease classification has to be settled based on the child's presenting complaint. But for any child who visited under-five OPD with presenting complaint of not taking any food, fluid or breast feeding, without any consideration of HMIS admission disease classification the child should be admitted to pediatric ward for parenteral feeding. Out of children presenting with this complaint (1051/1069) 98.32% were admitted to pediatric ward. This is an interesting and new knowledge that is hidden in under-five OPD data set.

Else if (unable to feed/drink=no and sunken eyes=no and WFA=LWFA) then the outcome of under-five OPD visit = admitted

- Another interesting rule discovered in this study is for the child able to feed or drink, first look the weight for age classification of the child. If it is low weight for age but not true for sunken eye then the outcome of under-five OPD visit is also admission. Therefore, without any consideration of HMIS admission disease classification, again for children who are able to feed or drink and whose weight is under low weight for age classification only by looking absence of sunken eyes the child is admitted to pediatric ward. Out of children presenting with this complaint (36/39) 92.31% were admitted to pediatric ward.

Else if (WFA=VLWA and restless/irritable=no) then the outcome of under-five OPD visit = admitted

- Regarding to the third interesting rule obtained in this study, again first look weight for age classification of the child. If the weight for age classification of the child is under very low weight for age but not in association with restless or irritable (i.e. the child's weight classification is not related with acute fluid loss from the body) then the outcome of under-five OPD visit is admission. Therefore, without any consideration of HMIS

admission disease classification, simply weight the child and look for the child's weight for age classification. If it is classified at under very low weight for age but the child is not irritable or restless then the child is admitted to pediatric ward only for its complaint of very low weight for age by excluding fluid loss. Out of children presenting with this complaint (354/360) 98.33% were admitted to pediatric ward

Else if (restless/irritable=yes) then the outcome of under-five OPD visit = not admitted.

- For the child present with the complaint of irritable/restless is true, admission is not recommended (6/6 100%). This is because this complaint is due to some dehydration. It is also usual to treat those children at ORS corner than admission.

Else if (WFA=AWFA and HMIS disease classification=malaria or TAPF or burn or severe pneumonia or meningitis or SAM or nephritis or hernia or cellulitis or others or intestinal obstruction or mastoditis or rectal prolapse or measles or tuberculosis or prematurity or anemia or pyomyocitis or RVI or RF or heart problems or meconium aspiration or DKA or tetanus or abscess or arthritis or hypothermia) then the outcome of under-five OPD visit = admitted

- For the HMIS disease classification of malaria (235/241, 97.51%), trauma accident, poison and fracture together, (130/169, 76.92%), burn (135/152, 88.82%), severe pneumonia (1964/1980, 99.19%), meningitis (30/30,100%), kwashiorkor (38/38, 100%), nephritis (51/51 100%), hernia (23/32 71.88%), cellulitis (65/82, 79.27%), intestinal obstruction (71/72, 98.67%), mastoditis (7/7, 100%), rectal prolapse (18/20, 90%), measles (17/17, 100%), anemia (20/20, 100%) , pyomyocitis (7/7, 100%), relapsing fever (1/1, 100%), angina (18/18, 100%), meconium aspiration (9/9, 100%), DKA (1/1, 100%), tetanus (6/6, 100%), abscess (30/30, 100%), arthritis (3/3, 100%), hypothermia (12/12, 100%), tuberculosis (171/172, 99.45%) and RVI (21/21, 100%) the child's weight for age classification is appropriate weight for age at admission. This is because except kwashiorkor, tuberculosis and RVI all other diagnoses are related with acute onset at admission. For the case of kwashiorkor, it is related with fluid retention in the body that resulting in weight gains of the child abnormally. Regarding to the RVI it is related with the positive result rather than the AIDS stage of the disease.

Else if (HMIS disease classification=Helminthiasis or UTI or skin infection or tonsillitis or foreign body or constipation or typhoid fever or conjunctivitis or dysentery or tumor or

lymphadenitis or ORTI or AFI or epistaxis or diarrhea with no dehydration or congenital deformity or unlikely bacterial infection or umbilical bleeding) then the outcome of under-five OPD visit = not admitted

- According to the decision rule in the output, admission is not recommended for the following HMIS disease classification at Nigist Eleni Mohammed Memorial Hospital: Helminthiasis (328/332, 98.8%), UTI (37/45, 82.22%), skin infection (485/537, 90.32%), tonsillitis (868/874, 99.31%), foreign body(44/65, 67.69%), constipation (50/55, 90.9%) typhoid fever (73/83, 87.95%), conjunctivitis (83/83, 100%), dysentery (50/51, 98.04%), tumor (11/11, 100%), lymphadenitis (94/94, 100%), ORTI (Common cold, bronchitis, asthma, croup) (99/106, 93.4%), AFI (141/178, 79.21%), epistaxis (7/8, 87.5%), diarrhea with no dehydration (3/3, 100%), congenital deformity (13/13, 100%) unlikely bacterial infection (18/18, 100%) umbilical bleeding (13/13, 100%). This is because, it is possible to manage all this cases at OPD level.

Else if (HMIS disease classification=diarrhea and skin pinch=slow) then the outcome of under-five OPD visit = admitted

Else if (skin pinch=very slow) then the outcome of under-five OPD visit = admitted

Else if (skin pinch=normal) then the outcome of under-five OPD visit = not admitted

Else if (HMIS disease classification=ORTI and fast breathing=yes) then the outcome of under-five OPD visit = admitted

Else if (fast breathing=no and vomiting=yes) then the outcome of under-five OPD visit = admitted

Else if (vomiting=no and diarrhea=yes) then the outcome of under-five OPD visit = not admitted

Else if (diarrhea=no) then the outcome of under-five OPD visit = admitted

Else if (HMIS disease classification=liver and GI problems and fever=yes) then the outcome of under-five OPD visit = admitted

Else if (fever=no and vomiting=no) then the outcome of under-five OPD visit = not admitted

Else if (vomiting=yes) then the outcome of under-five OPD visit = admitted

Else if (HMIS disease classification=genitalia problem and vomiting=yes) then the outcome of under-five OPD visit = admitted

Else if (vomiting=no) then the outcome of under-five OPD visit = not admitted

Else if (HMIS disease classification=infected wound and fever=yes) then the outcome of under-five OPD visit = admitted

Else if (fever=no and fast breathing=no) then the outcome of under-five OPD visit = not admitted

Else if (fast breathing yes) then the outcome of under-five OPD visit = admitted

Else if (HMIS disease classification=neonatal sepsis and vomiting=yes) then the outcome of under-five OPD visit = admitted

Else if (vomiting=no and fast breathing=yes) then the outcome of under-five OPD visit = admitted

Else if (fast breathing no and diarrhea=yes) then the outcome of under-five OPD visit = admitted

Else if (diarrhea=no) then the outcome of under-five OPD visit = not admitted

Else if (HMIS disease classification=pneumonia and stridor=yes) then the outcome of under-five OPD visit = admitted

Else if (stridor=no) then the outcome of under-five OPD visit = not admitted

Else if (HMIS disease classification=epilepsy and fever=no) then the outcome of under-five OPD visit = admitted

Else if (fever=yes) then the outcome of under-five OPD visit = not admitted

Else if (HMIS disease classification=seizure and convulsion=yes) then the outcome of under-five OPD visit = admitted

Else if (convulsion=no) then the outcome of under-five OPD visit = not admitted

Else if (HMIS disease classification=sepsis and fast breathing yes) then the outcome of under-five OPD visit = admitted

Else if (fast breathing=no) then the outcome of under-five OPD visit = not admitted

Else if (sunken eyes=yes and HMIS disease classifications=severe pneumonia or meningitis or diarrhea or foreign body or SAM or neonatal sepsis) then the outcome of under-five OPD visit = admitted

Else pneumonia or AFI =not admitted

4.7 Strengths and problems of the research

Some of the strengths of this study are:

- Data mining is applicable and yield good results if it is implemented in large volume of data at least greater than five thousand records. This research was done on more than fifteen thousand of records.
- The availability of high speed computer, that assists to get the outputs within short period of time.
- Using WEKA, which is freely available data mining tool for the study.
- The researcher has taken oral consent from Negist Eleni Mohammed Memoriaal hospital for the confidential issues related with the study.

Problems related with this research are:

- For mining, data are manipulated for the purpose of mining where data transformation, data reduction and filling missing values were done. There was no standard for the manipulation of the data during data preparation.
- Data was in log books that contains inconsistent records, missing values, noises etc. Vast majority of the researcher's time (80-90%) was spent on inputting data and preprocessing the data for mining.
- Lack of references in the area of this research topic for the comparison of findings.
- Shortage of time to cover all areas within specified period.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The underlying objective for undertaking this study was to apply data mining techniques on under-five dataset in developing a model. This model could support admission to pediatric ward, thereby enabling health service providers to make enhanced decisions in the effort to plan child health intervention programs.

Among the data mining classification techniques experimented, both decision tree and neural network showed comparative accuracy and performance for outcomes of under-five OPD visit. Models of decision tree and ANN are compared for the outcome of under-five OPD visit, in terms of accuracy, WTPR, WFPR, WROC and time required for building the models. The decision tree algorithm J48 has higher accuracy (94.77%), weighted true positive rate (94.7%), weighted false positive rate (5.3%), weighted receiver operating characteristics (0.99) and performs much faster than multilayer perceptron. In addition to that, models built by using neural network, are incomprehensible for a human and the extraction of business knowledge from it was found to be difficult. Hence, the decision tree approaches seems more appropriate to the users. Moreover, the decision tree algorithms have a simple feature which can be easily understandable by users.

According to interesting rules in J48, presenting complaint of not taking any food, fluid or breast feeding (98.32%), low weight for age without sunken eyes (92.31%) and very low weight for age but not in association with restless or irritable (98.33%) are among the cause of under-five children admission to pediatric ward without any consideration of health information management system admission disease classification.

In summary, encouraging results are obtained in classification tasks. Therefore, data mining technique is applicable on pediatric dataset in developing a model that support the discovery of the causes of under-five children admission to pediatric ward.

5.2 Recommendation

This study shows the potential applicability of data mining techniques in pediatric dataset in developing a classification model. Based on the study, the following recommendations are put forwarded for health managers, MOH, NGO’s, other stakeholders:

- Use of integrated dataset with other routes of child admissions such as emergency OPD, weekends and holidays data is also recommended.
- More research and development efforts need to be conducted to enable and explore the variety of data mining techniques that can be applied in pediatric dataset.
- Integration of data mining techniques into existing system and computerizing manual recording systems in database is a priority issue.
- Training is highly recommended for data handlers. Therefore, immediate managers of the organization, MOH, NGO's and other stakeholders must facilitate conditions for the overall improvement of data handling and storing.
- Besides computerizing the data, consulting experts on recording formats and information to be registered is also a crucial issue in improving the quality of health services.
- Implementation of the findings primarily in Nigist Eleni Mohammed Memorial hospital and other similar settings.
- The size of the dataset has an impact on data mining research. Especially proportional dataset will enhance the performance of the algorithms. Further researches can be conducted using large dataset.

References

1. Kantardzic M. *Data mining: concepts, models, methods and algorithms*. IEEE press; 2003.
2. Han J, Kamber M. *Data mining: concepts and techniques*. 2nd edition; Morgan Kaufmann publisher; 2006.
3. Hand D, Mannila H, Smyth P. *Principles of data mining*. The MIT press; 2001
4. UNICEF MDG, Reduce child mortality [On Line] 2011; Available from: http://www.unicef.org/mdg/childmortality.html14/10/2011 [Accessed 14th November 2011]
5. UNICEF, Statistical report on Ethiopia. [On Line] 2011; Available from http://www.unicef.org/infobycountry/ethiopia_statistics.html 14/10/2011 [Accessed 14th October, 2011]
6. United Nations, Millennium Development Goal report, New York, 2008.
7. Federal Ministry of Health, Maternal and child health package Addis Ababa, 2003.
8. Dangew M, Dante S, Daniel B. Analysis of admission to the pediatrics emergency ward of Tikur Anbesa Hospital. *EJHD* 2007; 21(1): 48-52.
9. Family health development FMOH, National strategy for child survival in Ethiopia, 2005
10. Witten IH. Frank E. *Data mining practical machine learning Tools and techniques.* 2nd ed. Elsiever Inc, USA: MorganKaufman publisher; 2005.
11. Cois KJ, Pedrycz W, Kurgan LA, Swiniarsski R. *Data Mining Methods a knowledge discovery approach*. New York U.S.A: Springer + Business media; 2007.
12. Nisbet R, Elder J, Miner G. *Hand book of statistical analysis and data mining applications.* Canada: Elsevier Inc; 2009.
13. Saxena S, Bottle A, Gilbert R, Sharland M, Increasing Short-Stay Unplanned Hospital Admissions among Children in England; Time Trends Analysis '97–06'. [On Line] 2009 e7484. doi:10.1371/journal.pone.0007484 PLoS ONE 4(10): Available from: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0007484 [Accessed 14th November 2011]

14 Rud Po. *Data mining Cook book: Modeling data for marketing, Risk, and customer relationship management.* John Wiley & Sons Inc. New York: Wiley computer publishing; 2001.

15 Cois KJ, Pedryez W, Kurgan LA, Swiniarski R. *Data Mining Methods a Knowledge Discovery Approach*. New York, U.S.A: Springer Science + Business media; 2007.

16 Tukuitonga CF, Robinson E. Hospital admission among pacific children. *NZ Med J* 2000; 113: 358-61.

17 Bianco A, Pileggi C, Trani F, Angelillo IT. Appropriateness of admission and days of stay in pediatrics ward of Italy. *Official Journal of American Academy of pediatrics* 2003; 112: 124.

18 Grenn C. Olsen primary General Hospital, Activity report [On Line] 2005; Available from: http://www.projectmercy.org/index.php?Itemid=52&id=16&option=com_content&task=view [Accessed 14th November 2011]

19 Amare D, Fasil T, Belaineh G. Determinants of under five mortality in Gilgel Gibe field research center, south west Ethiopia: *EJHD* 2007; 2(2):117-124.

20 Ponce J, Karahoca A. *Data mining and knowledge discovery in real life applications*. Published by In-The; 2009.

21 Two Crows Corporation. Introduction to data mining and knowledge discovery.[On line] 1999; Available from: http://www.twocrows.com/intro-dm.pdf [Accessed 18th November 2011]

22 Two Crows Corporation. Introduction to data mining and Knowledge discovery.[online] 2005; Available from:http://www.twocrows.com/intro-dm.pdf [Accessed 18th November 2011].

23 Ponce J, Karahoca A. *Data mining and knowledge discovery in real life application.* Crotia: Published by In.Teh; 2009.

24 Fayyad U, Piatetsky-Shapiro G, Smith P. From data mining to knowledge discovery in database. [Online] 1996; Available from: http://citeseer.nj.nec.com/fayyad96fromhtml [Accessed 16th November 2011].

25 Taft M, Krishnan R, Hornick M, Muhkin D, Tang G, Thomas S et.al *Oracle data mining concepts.* 10g Release2 (10.2), Oracle; 2005.

26 Berry M, Linoff G. *Data mining techniques for marketing, sales and customer relationship management.* 2nd ed. Indianapolis, Indiana: Wiley publishing, Inc; 2004.

27 Larose DT. *Discovering knowledge in data: An introduction to data mining.* Canada: A Jhon Wiley & Sons,Inc.,publication; 2005.

28 Dasu T, Johnson T. *Exploratory data mining and data cleaning.* Indianapolis Indiana: Wiley Pub. Inc; 2003.

29 Jain AK, Murty MN, Flynn PJ. 1999, Data clustering: *A review, ACM computing surveys* 1999: 31(3):264-323.

30 Olson DL, Delen D. *Advanced data mining techniques*. Berlin Heidelberg: Springer-verlag; 2008.

31 Mekasha A. Determinants of diarrheal diseases: A community based study in urban south west Ethiopia. *EAMJ* 2003; 80(2).

32 Girma, Berhane, Children who were vaccinated, breast fed and from low parity mothers lives longer: A community based case-control study in Jimma, Ethiopia [On Line] BMPC Public Health.2011; 11:197

33 Teshome B, Getahun Z, Taye G, Kogi-Makau W. Magnitude and determinants of stunting in children under-five years of age in food surplus regions in Ethiopia: The case of West Gojam Zone. *EJHD* 2009; 23(2): 98-106.

34 World Health Organization, Global Database on Child Growth and Malnutrition, 2006.

35 Mekonnen D. Infant and child mortality in Ethiopia: the role of socioeconomic, demographic and biological factors in the previous five years period of 2000-2005, 2011.

36 Bouma R, Scheitzer R, The impact of chronic child hood illness on family stress. [On Line] 1990 available from: http://www.ncbi.nlm.nih.gov/pubmed/12424981[Accessed 7th December 2011]

37 Zolten K. 1997, How parents can help their child with a chronic illness. Available from: http://www.parenting-ed.org/handout3/Specific%20Concerns%20and%20Problems/coping%20with%20chronic%20illness.htm[Accessed 7th December 2011]

38 Royal college of psychiatrics, Impact of chronic illness on parent's psychology. [On Line] 2004; Available from:

http://www.rupsych.ac.uk/mentalhealthandgrowingup/chronicphysicalillness.aspx [Accessed 14th November 2011]

39 B. D. Ripley and R. M. Ripley, Neural networks as statistical methods in survival analysis, in Artificial Neural Networks: Prospects for Medicine, R. Dybowski and V. Grant, Eds. Texas: Landes Biosciences, 1997.

Glossary

Admission – is an entry to the pediatric ward where under-five child will get treatment and care for some specified disease that needs special care.

Child – is developmental stage in human being from birth to puberty who seeks care of others.

Child mortality rate – is the number of deaths in children aged 1-4 years in a year per number of children in the same age in the same year.

Cause of admission – is an illness for that, under-five children will be admitted to pediatric ward where special care and treatment will be provided for specified disease.

Discharge – is arrangement of conditions for admitted child to go home when assumed to be the child's condition is improved or else died.

Length of stay – is period of hospitalization from date of admission to discharge in pediatric ward.

HMIS disease classification – is classification of disease that fulfills HMIS criteria for that the child get treatment or care in pediatric ward or under-five clinic

Morbidity – is the relative frequency of diseases occurrence on under-five children.

Mortality – the number of death that occurs at specific time in under-five children, from specific cause.

Under-five clinic – is a place where under- five children get treatment and care for their illnesses by medical person.

Under-five child- is a developmental stage in human being from birth until celebrating the 5^{th} year of age.

Under-five mortality rate – is number of deaths of children under five years in a year per number of live births in the same year.

Infant – is developmental stage in human being from birth to less than one year of age.

Infant mortality rate – is number of deaths among children under one year of age per 1000 live births in a year.

Neonate - is developmental stage in human being from birth to 28 days of age.

Neonatal mortality rate - is number of deaths of children under 28 days of age in a year per 1000 live births in the same year.

Pediatrics – the branch of medicine dealing with the care and development of children and with prevention and treatment of children's disease.

Pediatric ward – is a room where children under the age of 15 years get treatment and care for their illnesses

Annex A- Selected attributes for predicting cause of under-five children admission in under-five OPD

S.No	Attributes	Value it takes	Composition in number	Composition in %
1	Age in months	Less than 7.5	4849	30.64
		7.5 and above	10975	69.36
2	Sex	Male	9065	57.29
		Female	6759	42.81
3	Address	Hossana	9643	60.94
		Lemo	2923	18.47
		Misha	775	4.90
		Shashogo	318	2.01
		Analemo	348	2.20
		Soro	295	1.86
		Gibe	209	1.32
		KTzone	281	1.78
		Gombora	278	1.76
		Siltezone	217	1.37
		Duna	338	2.14
		Badewacho	26	0.16
		Guragezone	162	1.02
4	Weight in kg	Less than 1.7	24	0.15
		1.7-2.7	56	0.35
		2.7-5.95	1583	10.00
		5.95-10.25	7311	46.21
		Greater than 10.25	6850	43.29
5	Visit	Initial	15257	96.42
		Follow up	567	3.58
6	Cough	Yes	7639	48.27
		No	8185	51.73

7	Fever	Yes	4807	30.38
		No	11017	69.62
8	Diarrhea	Yes	3056	19.31
		No	12768	80.69
9	Vomiting	Yes	3147	19.89
		No	12677	80.11
10	Abdominal cramp/pain	Yes	569	3.60
		No	15255	96.40
11	Trauma/injury/accident	Yes	86	0.54
		No	15738	99.46
12	Others	Yes	2832	17.90
		No	12992	82.10
13	Unable to breast feed/drink	Yes	1079	6.82
		No	14738	93.18
14	Convulsion history/ convulsing now	Yes	238	1.50
		No	15585	98.50
15	Lethargic or unconscious	Yes	292	1.85
		No	15532	98.15
16	Fast breathing	Yes	7076	44.72
		No	8748	55.28
17	Stridor	Yes	587	3.71
		No	15237	96.29
18	Chest indrawing	Yes	2332	14.74
		No	13492	85.26
19	Bulged fontanelle	Yes	29	0.18
		No	15795	99.82
20	Restless/irritable	Yes	157	0.99
		No	15667	99.01
21	Blood in stool	Yes	91	0.58
		No	15733	99.42

22	Sunken eyes	Yes	667	4.22
		No	15157	95.78
23	Skin pinch	Normal	15087	95.34
		Slow	473	2.99
		Very slow	264	1.67
24	Weight for age	Appropriate weight for age	15319	96.85
		Low weight for age	443	2.80
		Very low weight for age	54	0.35
25	Oral trash	Yes	7	0.05
		No	15817	99.95
26	Immunization status	Completed	9211	58.21
		Up to date	5392	34.08
		Not started	1214	7.67
		Defaulted	7	0.04
27	HMIS disease classification	Malaria	462	2.92
		TAPF	131	0.83
		Tonsillitis	1679	10.62
		Helminthiasis	327	2.07
		Burn	142	0.90
		Severepneumonia	2500	15.80
		Pneumonia	4064	25.68
		UTI	36	0.23
		Skininfection	484	3.07
		Meningitis	151	1.04
		diarrhea	2758	17.43
		Foreignbody	44	0.28
		Constipation	49	0.31
		SAM	383	2.42
		Othersrespiratoryinfections	496	3.14
		Typhoidfever	72	0.46

		Nephritis	52	0.33
		Conjunctivitis	82	0.52
		Hernia	22	0.14
		Genitalproblems	87	0.55
		Infectedwound	108	0.68
		Cellulitis	64	0.41
		Liver and GIproblems	35	0.22
		Dysentery	49	0.31
		Tumor	11	0.07
		Neonatalsepsis	465	2.94
		Lymphadenitis	93	0.59
		Other not specified illnesses	50	0.32
		intestinalobstruction	73	0.46
		Epilepsy	19	0.12
		Mastoditis	7	0.04
		Rectalprolaps	17	0.12
		Measles	17	0.12
		AFI	146	0.92
		Seizure	41	0.26
		Tb	172	1.18
		Prematurity	71	0.46
		Anemia	41	0.26
		Epistaxis	7	0.04
		Pyomyocitis	7	0.04
		RVI	21	0.13
		RF	3	0.02
		Heartproblems	54	0.34
		meconiumaspiration	9	0.06
		DKA	7	0.04
		Tetanus	17	0.12

		Arthritis	3	0.02
		Abscess	32	0.20
		Sepsis	64	0.41
		Hypothermia	12	0.08
		congenitaldeformity	12	0.08
		unlikelybacterialinfection	17	0.12
		Umblicalbleeding	12	0.08
28	Outcome of under-five OPD visit	Admitted	5846	36.94
		Notadmitted	9978	63.06

=== Run information ===

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

Relation: 16selected

Instances: 11774

Attributes: 16

Fever, diarr., vom., D/una, D/con, D/let, fast, stridor, chest, bulfon, rest, sunk, skin, wfa, HMIS, class

Test mode: 10-fold cross-validation

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

J48 pruned tree

D/una = yes: admitted (1050.62/18.0)

D/una = no

| sunk = no

| | wfa = appropriateweightforage

| | | HMIS = malaria: admitted (234.17/6.17)

| | | HMIS = TAPF: admitted (129.09/40.09)

| | | HMIS = helminthiasis: notadmitted (327.24/4.0)

| | | HMIS = burn: admitted (134.1/18.1)

| | | HMIS = severepneuminia: admitted (1963.71/16.42)

| | | HMIS = UTI: notadmitted (36.03/8.0)

| | | HMIS = skininfection: notadmitted (484.35/52.0)

| | | HMIS = meningitis: admitted (30.02/0.02)

| | | HMIS = tonsilitis: notadmitted (867.63/6.0)

| | | HMIS = diarrhea

| | | | skin = normal: notadmitted (970.73/119.0)

| | | | skin = veryslow: admitted (2.0)

| | | | skin = slow: admitted (38.0/3.0)

| | | HMIS = foreignbody: notadmitted (43.03/21.0)

| | | HMIS = constipation: notadmitted (49.04/5.0)

| | | HMIS = SAM: admitted (38.03/0.03)

| | | HMIS = otherrespiratoryinfections

| | | | fast = no: notadmitted (324.16/40.0)
| | | | fast = yes
| | | | | vom. = yes: admitted (6.0)
| | | | | vom. = no
| | | | | | diarr. = no: admitted (50.12/17.12)
| | | | | | diarr. = yes: notadmitted (11.0/1.0)
| | | HMIS = typhoidfever: notadmitted (72.05/10.0)
| | | HMIS = nephritis: admitted (50.04/0.04)
| | | HMIS = conjunctivitis: notadmitted (82.06)
| | | HMIS = hernia: admitted (22.02/10.02)
| | | HMIS = genitalialproblems
| | | | vom. = yes: admitted (2.03/0.03)
| | | | vom. = no: notadmitted (85.04/19.0)
| | | HMIS = infectedwound
| | | | fever = yes: admitted (3.03/0.03)
| | | | fever = no
| | | | | fast = no: notadmitted (103.04/25.0)
| | | | | fast = yes: admitted (2.0)
| | | HMIS = cellulitis: admitted (64.05/18.05)
| | | HMIS = liverandGIproblems
| | | | fever = yes: admitted (15.01/0.01)
| | | | fever = no
| | | | | vom. = yes: admitted (2.01/0.01)
| | | | | vom. = no: notadmitted (15.0)
| | | HMIS = pneumonia
| | | | stridor = no: notadmitted (2007.34/7.0)
| | | | stridor = yes: admitted (14.0)
| | | HMIS = dysentry: notadmitted (49.04/1.0)
| | | HMIS = tumor: notadmitted (11.01)
| | | HMIS = neonatalsepsis
| | | | vom. = yes: admitted (212.13/0.13)

| | | | vom. = no
| | | | | fast = no
| | | | | | diarr. = no: notadmitted (180.04/56.0)
| | | | | | diarr. = yes: admitted (3.0)
| | | | | fast = yes: admitted (26.13/0.13)
| | | HMIS = lymphadinitis: notadmitted (93.07)
| | | HMIS = others: admitted (45.03/2.03)
| | | HMIS = intestinalobstruction: admitted (70.05/1.05)
| | | HMIS = epilepsy
| | | | fever = yes: notadmitted (9.01/3.0)
| | | | fever = no: admitted (10.01/0.01)
| | | HMIS = mastoditis: admitted (7.01/0.01)
| | | HMIS = rectalprolaps: admitted (17.01/3.01)
| | | HMIS = measle: admitted (17.01/0.01)
| | | HMIS = otherrespiratoryinfectionss: notadmitted (98.07/7.0)
| | | HMIS = AFI: notadmitted (140.1/37.0)
| | | HMIS = seizure
| | | | D/con = no: notadmitted (10.02/1.0)
| | | | D/con = yes: admitted (15.0)
| | | HMIS = tb: admitted (171.12/1.12)
| | | HMIS = prematurity: admitted (16.01/0.01)
| | | HMIS = anemia: admitted (20.01/0.01)
| | | HMIS = epistaxis: notadmitted (7.01/1.0)
| | | HMIS = pyomyocitis: admitted (7.01/0.01)
| | | HMIS = RVI: admitted (21.02/0.02)
| | | HMIS = RF: admitted (1.0/0.0)
| | | HMIS = heartproblems: admitted (18.01/0.01)
| | | HMIS = meconiumaspiration: admitted (9.01/0.01)
| | | HMIS = diarrhea: notadmitted (3.0)
| | | HMIS = DKA: admitted (1.0/0.0)
| | | HMIS = tetanous: admitted (6.0/0.0)

| | | HMIS = arthritis: admitted (3.0/0.0)
| | | HMIS = abscess: admitted (30.02/0.02)
| | | HMIS = sepsis
| | | | fast = no: notadmitted (32.02/1.0)
| | | | fast = yes: admitted (18.02/7.02)
| | | HMIS = hypothersmia: admitted (12.01/0.01)
| | | HMIS = congenitaldeformity: notadmitted (12.01)
| | | HMIS = unlikelybacterialinfection: notadmitted (17.01)
| | | HMIS = umblicalbleeding: notadmitted (12.01)
| | wfa = verylowweightforage
| | | rest = no: admitted (354.18/6.11)
| | | rest = yes: notadmitted (6.0)
| | wfa = lowweightforage: admitted (35.02/4.01)
| sunk = yes
| | HMIS = severepneuminia: admitted (16.0/8.0)
| | HMIS = meningitis: admitted (1.0)
| | HMIS = diarrhea: admitted (535.0/12.0)
| | HMIS = foreignbody: admitted (1.0)
| | HMIS = SAM: admitted (51.0)
| | HMIS = pneumonia: notadmitted (11.0)
| | HMIS = neonatalsepsis: admitted (1.0)
| | HMIS = AFI: notadmitted (6.0/1.0)

Number of Leaves : 131

Size of the tree : 153

Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	11111	94.3689 %
Incorrectly Classified Instances	663	5.6311 %
Kappa statistic	0.8873	
Mean absolute error	0.0874	

Root mean squared error 0.2121

Relative absolute error 17.4761 %

Root relative squared error 42.4315 %

Coverage of cases (0.95 level) 99.1507 %

Mean rel. region size (0.95 level) 63.2283 %

Total Number of Instances 11774

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.922	0.035	0.963	0.922	0.942	0.979	admitted
0.965	0.078	0.926	0.965	0.945	0.979	notadmitted
Weighted Avg. 0.944	0.057	0.944	0.944	0.944	0.979	

=== Confusion Matrix ===

a	b	<-- classified as
5376	457	a = admitted
206	5735	b = notadmitted