Addis Ababa University
School of Graduate studies
School of Information Science

Application of Data Mining Techniques to Predict Children Datasets The Case of Love for Children Organization

Gebremedhin Gebreyohans

June, 2012

Addis Ababa University
School of Graduate studies
School of Information Science

Application of Data Mining Techniques to Predict Children Datasets The Case of Love for Children Organization

A Thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the degree of Master of Science in Information Science

By
Gebremedhin Gebreyohans

June, 2012

Addis Ababa University
School of Graduate studies
School of Information Science

Application of Data Mining Techniques to Predict Children Datasets
The Case of Love for Children Organization

By
Gebremedhin Gebreyohans

Name and signature of Member of the Examining Board

Name Title Signature Date

_____________________ Chairperson, ____________ __________

_____________________ Advisor (s), _____________ __________

_____________________ Examiner, ______________ __________

DEDICATION

This research is dedicated to the “almighty of God” that is always there for me in all the hard time and challenges of my life.

Acknowledge

I would like to thank my advisor Ato Getachew Jemaneh, for his constructive suggestions and overall guidance.

Finally, I want to thank Selam Nguse for understanding and support in my everyday life.

Last but not least, I would like to thank the Love for Children Organization staff, W/o Lemlem Tikuye in particular, without whose good will to supply me with the data; I would not have accomplished this research output.

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
ARFF:	Attribute Relation File Format
CRISP-DM:	Cross Industry Standard Process for Data Mining
CSV:	Comma Separated Value
DM:	Data Mining
GUI:	Graphical User Interface
HIV:	Human Immunodeficiency Virus
KDD:	Knowledge Discovery in Databases
LCO:	Love for children Organization
NGO:	Non-Governmental Organization
OVC:	Orphans and Vulnerable Children
SEMMA:	Sample Explore Modify Model Assess
UNICEF:	United Nations Children's Fund
WEKA:	Waikato Environment for Knowledge Analysis

Table of Contents

LIST OF TABLES

LIST OF FIGURES

APPENDICES

ABSTRACT

This research describes and evaluates classification of children to four classes as orphan, single orphan, vulnerable and safe that will help for the organizations to donors from outside of the organization and to get full information about each child for internal purpose of the organization. To classify this three algorithms of classification techniques are used which are Decision tree, Bayesian learning and Neuron Network algorithm have been explored within the framework of KDD data mining model has been used and The data have been analyzed and interpreted using the WEKA 3.7.4 version software. Data are collected, cleaned, transformed and integrated for experimenting with the classification model. The final dataset consists of 17044 records have been experimented and evaluated against their performances.

The collection of data set are experimented with the 10-fold cross-validation and splitting the datasets in to 70% for training and 30% for testing, 66% for training and 44% for testing as well as 50% for training and 50% for testing is used. Finally a comparison of decision tree, Bayesian network and neural network model in terms of the overall classification accuracy and their advantage is made. As a result, Decision Tree is selected because it gives better results than Bayesian learning and better advantage over Neuron Network so due to more advantages of decision tree over the others: The accuracy of these algorithms are by Decision Tree C4.5, and Naïve Bayes, Multilayer Perception, 98.83%, 98.32%, 98.86% respectively. This shows that the Decision Tree classifier performed better than the other on a specific children's dataset.

Finally, in overall decision tree is selected as a model for the organizations children classification.

Chapter one

Introduction

1.1. Background

Data mining is a new kind of business information processing technology, which can extract interesting patterns or knowledge implicated in a large number of incomplete, noisy, and ambiguous data that people do not know in advances but with potential application (Han and Kamber 2001). In addition to this Han and Kamber stated that, Nowadays data can be stored in many different types. The steady and amazing growth of computers and information technology even exposed the availability of data on different location with various formats. The abundance of data, coupled with the need for powerful data analysis tools has been described as data rich but information poor society.

However, according to prominent researchers in the field (Fayyd et al., 1996), despite their popular interchangeable usage, there is significant difference between data mining and KDD. While KDD refers to the whole process of changing low level data into high level knowledge, data mining constitutes one of the phases in this process, namely "the application of specific algorithms for extracting patterns from data" (Fayyd et al., 1996).

In the Information Technology era information plays vital role in every sphere of the human aspects. Thus, to efficiently inspire information, it is very important to generate information from massive collection of data. The data can range from simple numerical figures and text documents to more complex information such as spatial data, multimedia data, and hypertext documents (Deshpande and Thakare, 2010). However, the huge size of these data sources makes it impossible for a human experts to come up with interesting information or patterns that will help in the proactive decision making process. Therefore, to take complete advantage of data; the data retrieval is not enough. It requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. This tool is data mining (Deshpande and Thakare, 2010).

Data mining techniques can be applied to a wide variety of data repositories including databases, data warehouses, spatial data, multimedia data, Internet or Web-based data and complex objects (Lori, 2006). Advances in computer hardware and data mining software have made data mining systems accessible and affordable to many businesses. Its application areas can be educational, banking, medicine, financial, criminology, and many others (Deshpande and Thakare, 2010). Hence, it is not surprising that data mining has gained widespread attention and increasing popularity in governmental and Nongovernmental organizations like helping childes those in need of helping in recent years. Different organizations have been adapting data mining to their own environment, in their own way. As it is the case in other sectors of the society, the NGO community also depends on the availability of information for decision-making purpose. NGO is a broad term encompassing a wide array of diverse organizations. According to the World Bank, the term NGO refers to "private organizations that pursue activities to relieve suffering, promote the interests of the poor, protect the environment, provide basic social services or undertake community development" (World Bank, 2001).

Now days, there are a lot of children who are in need of help. There are a number of organizations who are providing help for the needy children. Including Love for Children Organization is non-governmental Organization. Currently, the organization is performing child selection by consulting psychologists and social workers to decide whether the child needs help or not. This creates tremendous human involvement, ambiguity, and unnecessary cost of the consulting personnel to the organization. This in turn results in the possibility of allocating fund for non needy children of the organization. Hence, an automated system can help facilitate and simplify the process of identifying and selecting children who are in need of help.

In this paper, the researcher explained application of data mining to orphan and vulnerable children and the potential of data mining becoming very important and wide range of companies around the globe has deployed successful application of data mining (Thearling, 2003). However only few researches have been conducted at the School of Information Science, Addis Ababa University.

The first attempt was made by Shegaw in 2002 that was undertaken a research on children related data mining application that is on predicting child mortality patterns. Similarly, Woldekidan in 2003 has done similar research on street children Ethiopia and Helen in 2003 also conducted on the application of data mining technology to identify significant patterns in census or survey data: in Ethiopian 2001 child labor survey. Hence, this research is a continuation of the data mining researches carried out so far, however, with different area of application, which is data mining application for vulnerable and orphan children in the Love for Children organization (LCO).

According to the broacher of the organization Love for Children is nongovernmental organization established in 1999 by acquiring its legal personality from ministry of justices. Since its establishment LCO has engaged itself in various programs. The program that LCO is operating have brought about significant impact on in the lives of children as well as the community at large and the main objective of the organization is fight for ensuring children's well being and is working to bring about lasting improvement upon their lives though strengthening institutions to support them and take actions with a vision of see Ethiopia being a place where children can grow up in a favorable environment and with promising opportunities to become self reliant productive member of the society.

1.2. Statement of the Problem

For data mining, huge amount of data is required to generate information the data can be simple numerical figures and text documents, or more complex information like spatial data, multimedia data and hyper text documents. To take complete advantage of data stored in files, data bases, and other repositories, data retrieval is simply not enough. It requires a technique or powerful tool for analysis and interpretation of such data and that could help in decision-making. This technique or tool is data mining. Data mining is extraction of hidden predictive and descriptive information from large data bases or huge data. It is a power full technology with great potential to help organizations to focus on the most important information in their data ware houses.

The topic chosen addresses techniques for classifying children data. This system when incorporated with visualization techniques can be used as a useful tool for decision making purpose for the organization. Such that classifying children as orphan or vulnerable based on the children's health condition and status of parents and to give information for the organization provides funds for children from different segments of the community who are assumed to be needy.

So, this study is aimed at data collection, preparation, updating, classification and visualization and analyzing the data items in children data and extract useful patterns from the given data. For example, classifying children as orphan or vulnerable based on the children's health condition and status of parents. To this end the study explores to analysis the questions. There are researches that have been done related to children using data mining techniques in Ethiopia organizations context by Shegaw (2002), Woldekidan (2003) and Helen (2003).

This research complements with the research work mentioned above. But it differs in the area of application, variables used for model building, the theme it contains, and data mining methods.

Thus, in the present research an attempt is made to address the following research questions:-

- What knowledge is suitable for Data mining in classification of children to acquire and modeled data mining algorithms?
- How huge amount of data can be represented in data mining algorithms using classification for future use in decision making.
- Can data mining helps to solve the current problem of child classification in Love for Children Organization?

1.3. Objective of the Study

The study will have the following general and specific objectives respectively.

1.3.1. General Objective

The general objective of this study is to develop a model that helps LCO identify different patterns of children data through the application of classification techniques so that identification of different classes of children can be automated (i.e., to decide whether a child is vulnerable, orphan, single orphan or safe based on the given data).

1.3.2. Specific Objectives

In order to achieve the general objectives, the following specific objectives are stated.

- To acquire domain knowledge using interview and document analysis, i.e., to extract the data, clean and transform the data into the format suitable for mining.
- Reviewing different literatures that can support for the study.
- Preprocessing and preparing the raw data into a suitable dataset for experiment.
- Exploring and selecting appropriate data mining tool, techniques and functionalities.
- Building models using the preprocessed dataset.
- To evaluate (test) the model and visualize the result as well as
- Forward recommendation for further research.

1.4. Research Methodology

Methodology means the steps or procedures that the researcher follow to achieve the objectives that are stated in its specific and general objectives and it is a road map that shows the direction how the research is going to be done to reach the end. In order to achieve the proposed objectives the various steps followed and tasks accomplished the following methodologies have been used.

1.4.1. Understanding the Problem Domain

In this research, to solve the problems mentioned above, the researcher adopted the KDD process model have been used. This is because it provides basic knowledge consolidation, incorporating this knowledge into the system for effective management of data, finding patterns and relationships. KDD is widely used data mining model that includes problem domain understanding, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results and provides popular application (Roshan, 2011)

In order to define and analyze the business problem properly, the primary data was collected by interviewing concerned officers (Experts) in the organization. The offices that the researcher has been conducted the survey are the database of the Love for children organization. Then based on the information obtained from these attempts, the overall children classification process has been done. The model employed in this research is KDD process which consists of five phases (data selection, preprocessing, transformation, data mining and interpretation/evaluation).

1.4.2. Understanding the Data

The data to be mined was collected and arranged into a new database to make it suitable for the experiment and for the selected data mining tool. That means a new database was prepared by analyzing the collected data using data preprocessing tasks like data cleaning, data reduction and data transformation.

Among data mining techniques classification have been used for the interpretation of the data that enables to identify among different collection of children data.

1.4.3. Implementation Tool

Even though there are several data mining tools that may fulfill the objectives, techniques and tasks the researcher used for this, Weka 3.7.4 data mining and knowledge discovery tool. Weka, a machine-learning algorithm in Java, was adopted for undertaking the experiment. Weka constitutes several machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. Weka is also well suited for developing new machine learning schemes. Weka is open source software.

1.4.4. Data Preparation

At this step, the dataset on which the algorithm is to be run on is constructed. For one thing, the data, which was stored in 12 different tables and later, merged into one large table.

For another, since the convenient algorithm selected for this study is Decision tree (C4.5), Naïve Bayes and Neural Network algorithm. Which is the selected data mining algorithm, is often used in situations where attributes are nominal, the data was transformed to fit this algorithm. Other tasks handling missing data, and handling noise and errors.

1.4.5. Data Mining

The actual data mining took place at this stage. Based on the identified goals and the assessment of the available data, appropriate mining algorithm was chosen and run on the prepared data.

As one can tell from the objectives of the research, gaining knowledge, discovering pattern within the Children Data base, is certainly the purpose of this study. Having this purpose in mind, supervised learning technique, namely classification technique is adopted.

1.4.6. Literature Review

Various data sources will be used to gather the required information. These sources can be classified as primary and secondary sources. Primary sources such as the structured interview with domain experts to acquire the necessary information and secondary sources such as books, articles, journals and the internet will be referred to investigate the principles/theories of the various approaches, techniques and tools that will be employed in different areas.

1.4.7. Evaluation and Interpretation Mechanism

The performance of the system will be evaluated by comparing the decision suggested by the system against domain expert. In addition the effectiveness of the retrieval is tested using recall and precision.

1.5. Scope and Limitation of the Research

As a general schema there are different mechanisms to apply data mining techniques Thus, the scope of this research is limited to the Classification of huge data and producing use full patterns. This research focuses on supervised data mining techniques, which is classification were the chosen techniques. But unsupervised learning schemes such as clustering and association rule mining schemes were not used.

The prototype handles only classification of children to four categories like orphan, single orphan, Vulnerable and safe that provides advices in that field of study.

1.6. Significance of the Study

The significance of the study lies in making a contribution on the concept of Data mining in Love for Children Organization. It also promotes system analysis and programmers to view the opportunities and participate in designing and implementing Data mining using different data mining algorithms like classification for curbing the problems that exist in Love for Children Organization. Furthermore; it will create a general awareness among Love for Children Organization members and any interested on how Data mining can implement using classification techniques has a direct impact on children classification problems. Finally, Computer program that represents and reasons with knowledge of some specialist subject with a view to solving problems or giving advice.

1.7. Organization of the Study

This thesis is organized into five chapters. The first chapter is an introductory part, which discusses the background to the research work, problem area leading to this research, the general and specific objectives to attain in the research and the methodology to be followed. The second chapter mainly revolves around the Literature reviewed to know and write about meaning and importance of data mining, steps involved in data mining process and about different types of data mining tasks and algorithms. A detailed discussion of the algorithm to be utilized in attaining the goal of the data-mining task is also made. Chapter three explains the methods, the classification techniques, used in this research. The fourth chapter provides discussions about the different data mining steps that were undertaken in this research work. This includes data collection, data selection, preparation, model building and evaluating and interpreting results obtained from J48 decision tree, Bayesian network and neural network.

The last chapter is devoted for the final conclusions and recommendations based on the research findings.

Chapter Two

Literature Review

2.1. Data Mining Concepts

2.1.1. Introduction

This chapter presents review of related literatures on data mining topics. This includes what data mining is, different phases evolved on data mining, varieties of models, decision tree and various activities involved in it. This enables to easily understand the focus of the study and application of data mining for the problem domain.

2.1.2. Data Mining

Data mining is a powerful tool because it can provide you with relevant information that you can use to your own advantage. When you have the right knowledge, all you will need to do is apply it in the right manner, and you will be able to benefit. It is relatively easy to get information these days. But it is not so easy to get relevant information that can help you achieve a desired goal (Nanda, 2010).

Mining data and knowledge from large database has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues (Elkan, 1997). Data store in a database where as knowledge store in a knowledge base (Roshan, 2011). Across a wide diversity of fields, data are being collected and accumulated at a remarkable speed. As Roshan (2011) discussed, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.

Data mining refers to the process of finding interesting patterns in a large database that are not explicitly part of the data. The extracted interesting patterns can be used to tell us something new and to make predictions and description (Lori, 2006).It is a young interdisciplinary field drawing from areas such as: database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, neural networks, pattern recognition, spatial data analysis, and many more. (Han and Kamber, 2006)

2.1.3. Data Mining and Knowledge Discovery in Databases (KDD)

Knowledge discovery and data mining (KDD)—the rapidly growing interdisciplinary field which merges together database management, statistics, machine learning and related areas—aims at extracting useful knowledge from large collections of data. Knowledge Discovery is a process that seeks new knowledge about an application domain. It consists of many steps, one of them being Data Mining (DM), each aimed at completion of a particular discovery task, and accomplished by the application of a discovery method (Klosgen and Zytkow, 2002). Data mining is the search for relationships and global patterns that exist in large databases but are hidden among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database (Fayyad et al., 1996). As time passed, the amount of data in many systems, organizations and business agents grew to larger than terabyte size, and could no longer be maintained manually. KDD used to overcome the problem arises to manage data properly and to become successful in business activities which leads to discovering underlying patterns in data is considered essential. As a result, several software tools and techniques were developed to discover hidden data and make assumptions, which formed a part of artificial intelligence. The term KDD process refers to the whole process of changing low level data into high level knowledge which is automated discovery of patterns and relationships in large databases and data mining is one of the core steps in the KDD process.

The KDD process is interactive and highly iterative, user involved, multistep process; comprising a number of phases requiring the user to make several decisions KDD employs methods from various fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization (Fayyad et al., 1996). It is said to employ a broader model view than statistics and strives to automate the process of data analysis, including the art of hypothesis generation. Generally KDD can simply be defined as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad et al., 1996). The goal of knowledge discovery in data bases (KDD) and data mining (DM) is to find interesting patterns and/or models that exist in databases but are hidden among the volumes of data.

Data mining (DM) and knowledge discovery in databases (KDD) refer to analysis of huge digital data sets. (Adrians and Zantinge, 1996), define , “data mining is the analysis of (often large) observational data sets to fined unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The need for data mining arises from the huge digital data repositories’’.

Knowledge discovery in databases encompasses all the processes, both automated and none automated, that enhance or enable the exploration of databases, large and small, to extract potential knowledge. The most commonly referenced component of these processes has been data mining which involves activities oriented toward identifying patterns or models in data representation, classification, semantics, rules application, and so on (Fayyad et al., 1996).

Organizations are increasingly storing large amounts of data generated during their operating activities. Patterns that indicate the effectiveness of the various business processes are usually buried within this historical data. A recently utilized analysis method, data-mining, has the ability to discover patterns stored within historical data and is now considered a catalyst for enhancing business process by avoiding failure patterns and exploiting success patterns. It has been estimated that the quantity of data in the world roughly doubles every year, while the amount of meaningful information decreases rapidly (Hand, et.al.2001). Properly analyzing data and detecting these patterns is therefore of great importance to businesses.

Several data mining techniques have been developed over the last decade. Generally, the data mining techniques can be categorized in four categories, depending on their functionality: classification, clustering, numeric prediction, and association rules. The main difference between the different techniques is in the way they extract information (algorithms and methods used) and how results (knowledge discovery/rules) are expressed.

2.1.4. Knowledge Discovery Process

KDD is the "non trivial extraction of implicit, previously unknown and potential useful information from data", method to digest/find hidden information in database, and which convert unknown / hidden pattern into useful, understandable and informative way (Nanda, 2010)

KDD is the process of discovering useful knowledge from a collection of data. It concerns the knowledge discovery process applied to databases (Klosgen and Zytkow, 2002). It is also defined as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996).

Here Fayyad et al (1996) present their (necessarily subjective) perspective of a unifying process centric framework for KDD. The goal is to provide an overview of the variety of activities in this multidisciplinary field and how they fit together. KDD Process defined by Fayyad et al (1996) as: The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The term pattern goes beyond its traditional sense to include models or structure in data. In this definition, data comprises a set of facts (e.g., cases in a database), and pattern is an expression in some language describing a subset of the data (or a model applicable to that subset). The term process implies there are many steps involving data preparation; search for patterns, knowledge evaluation, and refinement all repeated in multiple iterations. The process is assumed to be nontrivial in that it goes beyond computing closed-form quantities; that is, it must involve search for structure, models, patterns, or parameters. The discovered patterns should be valid for new data with some degree of certainty. Fayyad et al (1996) also want patterns to be novel (at least to the system and preferably to the user) and potentially useful for the user or task. Finally, the patterns should be understandable if not immediately, then after some post processing.

Fayyad et al (1996) can define quantitative measures for evaluating extracted patterns. In many cases, it is possible to define measures of certainty (e.g., estimated classification accuracy) or utility (e.g., gain, perhaps in dollars saved due to better predictions or speed-up in a system's response time). Such notions as novelty and understandability are much more subjective. In certain contexts, understandability can be estimated through simplicity (e.g., number of bits needed to describe a pattern). An important notion, called interestingness, is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity. Interestingness functions can be explicitly defined or can be manifested implicitly through an ordering placed by the KDD system on the discovered patterns or models.

As described by Fayyad et al (1996) the KDD process consists of five steps:

- ✓ **data selection** – having two subcomponents: (a) developing an understanding of the application domain and (b) creating a target dataset from the universe of available data;
- ✓ **preprocessing** – including data cleaning (such as dealing with missing data or errors) and deciding on methods for modeling information, accounting for noise, or dealing with change over time;
- ✓ **transformation** – using methods such as dimensionality reduction to reduce data complexity by reducing the effective number of variables under consideration;
- ✓ **data mining** – having three subcomponents: (a) choosing the data mining task (e.g., classification, clustering, summarization), (b) choosing the algorithms to be used in searching for patterns, (c) and the actual search for patterns (applying the algorithms);
- ✓ **interpretation/evaluation** – having two subcomponents: (a) interpretation of mined patterns (potentially leading to a repeat of earlier steps), and (b) consolidating discovered knowledge, which can include summarization and reporting as well as incorporating the knowledge in a performance system.

Although the list above might suggest a linear process, KDD in practice is anything but linear. (Brachman et al, 1996), for example, contend that "knowledge discovery is a knowledge-intensive task consisting of complex interactions, protracted over time, between a human and a (large) database, possibly supported by a heterogeneous suite of tools." Fayyad et al (1996) emphasize both the interactive and iterative nature of KDD, that humans make many decisions and that the various steps and methods within them are repeated frequently as knowledge is being refined – thus our contention above that KDD is really about knowledge construction
Rather than discovery.

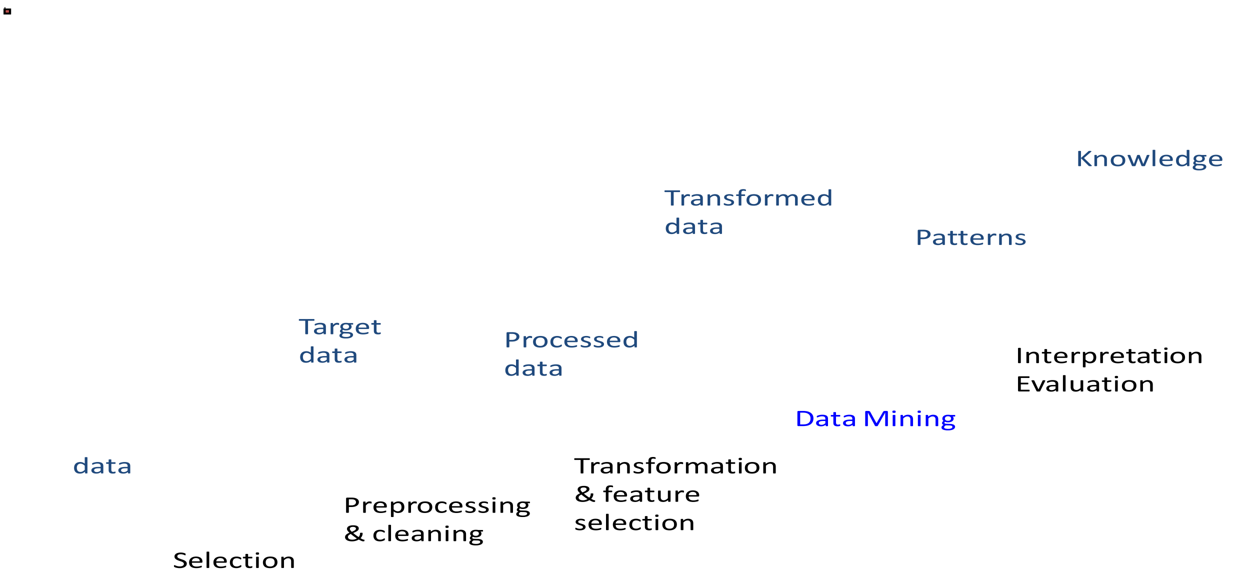


Figure 2.1: KDD Process: Adapted from: Fayyad et al (1996), "From Knowledge Discovery to Data Mining"

The KDD process will be classified in to sub major phases of KDD process i.e. per and post processing to control the activities and task parallel for find effective and efficient knowledge from unseen patterns using different data mining techniques and tools iteratively. Pre-processing of KDD process basic tasks are: learning the domain, creating a datasets, data cleaning, integration and transformation, data reduction and projection and choosing the data mining task. After completing Data Mining, we visualized the results and called post-processing. A possible post-processing methodology includes: find all potentially interesting patterns according to some rather loose criteria; provide flexible methods for iteratively and interactively creating different views of the discovered patterns, use of discovered Knowledge to find required patterns.

Data mining is a step in the KDD process consisting of an enumeration of patterns (or models) over the data, subject to some acceptable computational-efficiency limitations. Since the patterns enumerable over any finite dataset are potentially infinite, and because the enumeration of patterns involves some form of search in a large space, computational constraints place severe limits on the subspace that can be explored by a data mining algorithm.

As discussed by Fayyad et al (1996) in particular, data mining algorithms consist largely of some specific mix of three components:

- ✓ The model. There are two relevant factors: the function of the model (e.g., classification and clustering) and the representational form of the model (e.g., a linear function of multiple variables and a Gaussian probability density function). A model contains parameters that are to be determined from the data.
- ✓ The preference criterion. A basis for preference of one model or set of parameters over another, depending on the given data. The criterion is usually some form of goodness-of-fit function of the model to the data, perhaps tempered by a smoothing term to avoid over fitting, or generating a model with too many degrees of freedom to be constrained by the given data.
- ✓ The search algorithm. The specification of an algorithm for finding particular models and parameters, given data, a model (or family of models), and a preference criterion.

A particular data mining algorithm is usually an instantiation of the model/preference/search components (e.g., a classification model based on a decision tree representation, model preference based on data likelihood, determined by greedy search using a particular heuristic (Fayyad et al., 1996). Algorithms often differ largely in terms of the model representation (e.g., linear and hierarchical), and model preference or search methods are often similar across different algorithms.

2.1.5. Origin of KDD Model

A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) under data mining.

The phrase knowledge discovery in databases was introduced during the first KDD workshop in 1989 to stress that knowledge is the ultimate output of a data-driven discovery. KDD was then widely accepted in artificial intelligence and machine-learning fields by modeling of real-world phenomena (Lori, 2006). Knowledge Discovery in Databases (KDD) is the process of extracting novel information and knowledge from large databases. This process consists of many interacting stages performing specific data manipulation and transformation operations with an information flow from one stage onto the next (and often back into previous stages). The process can be very complex and may exhibit much variety in the context of the variety tasks undertaken within KDD (Matheus, et, all, 1993).

2.1.6. Why is KDD so Popular ?

According to Qi Luo (2008), large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses. Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in 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. For example, the marketing database of a consumer goods company may yield knowledge of correlations between sales of certain items and certain demographic groupings. This knowledge can be used to introduce new targeted marketing campaigns with predictable financial return relative to unfocused campaigns. Databases are often a dormant potential resource that, tapped, can yield substantial benefits.

The value of storing volumes of data depends on our ability to extract useful reports, spot interesting events and trends, support decisions and policies based on statistical analysis and inference and exploit the data to achieve business, operational, or scientific goals (Lori, 2006).

2.1.7. Tasks of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery (Dunham, 2003).

The goal of data mining is to produce new knowledge that the user can act upon. It does this by building a model of the real world based on data collected from a variety of sources which may include corporate transactions, customer histories and demographic information, process control data, and relevant external databases such as credit bureau information or weather data. The result of the model building is a description of patterns and relationships in the data that can be confidently used for prediction (Dunham and Sridhar, 2006).

According to Larose and Fayyad… et al (1996) data mining models are of two types such as descriptive and predictive. In other word the tasks of data mining can be modeled as either Predictive or Descriptive in nature (Dunham, 2003).

A Predictive model makes a prediction about values of data using known results found from different data while the Descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties.

Predictive model data mining tasks include classification, prediction, regression and time series analysis. The Descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis Selection of the modeling techniques is based upon the data mining objective. Modeling is an iterative process - different for supervised and unsupervised learning.(Fadzilah et. Al,. nd).

Dunham (2006) suggested the following conceptual hierarchy of Data mining task and models

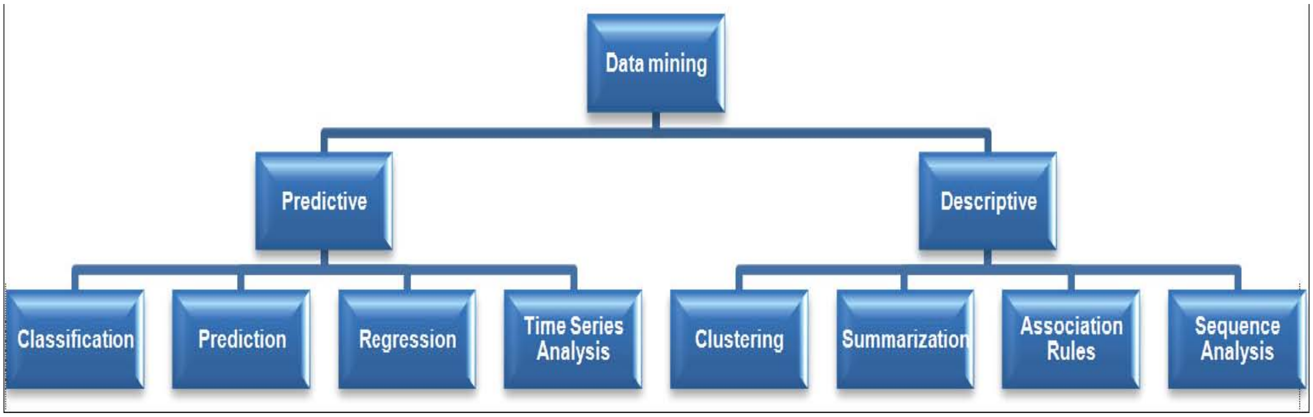


Figure 2.2 Data mining tasks and methods (Sources Dunham)

2.1.7.1. Predictive Modeling

Data mining is usually described as collecting information from recorded data, hence the name, mining. Predictive modeling is the event in which a model is made or chosen to accurately predict an outcome. Businesses have grown and science is a lot more complex, therefore data mining requires more computerized method of doing this. Software and computer applications have made data mining seem an effortless job, because it does not require a direct hands-on approach (Dunham and Sridhar, 2006). Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state.

Predictive modeling is the process by which a model is created or chosen to try to best predict the probability of an outcome and In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data (Han and Kamber, 2001).It includes mainly Classification, Regression, and Time series analysis. It uses some variables to forecast unknown or future values of other variables while the descriptive ones find patterns to describe the data.

Classification:-Among the Predictive models, Classification is probably the best understood of all data mining approaches. Three common characteristics of classification tasks are (Dunham, 2003).

- Learning is supervised
- The dependent variable is categorical
- The model built is able to assign new data to one of a set of well-defined classes.

Regression:-The regression involves the learning of function that map data item to real valued prediction variable.

Time series analysis:-In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. It includes mainly Clustering, Summarization, Association rule and Sequence discovery. Many of the data mining applications are aimed to predict the future state of the data. The predictive model makes prediction about unknown data values by using the known values.

2.1.7.2. Descriptive Modeling

Descriptive data mining model is the unsupervised learning functions. These functions do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness, etc. of the data. It presents the main features of the data or a summary of the data. Data randomly generated from a "good" descriptive model will have the same characteristics as the real data (Fadzilah et. Al, nd).

A descriptive modeling technique, such as Summarization, Association Rules, Sequence Analysis and clustering which produces classes (or categories), which are not known in advance.

Summarization: - maps data into subsets with associated simple descriptions. Basic statistics such as Mean, Standard Deviation, Variance, Mode and Median can be used as Summarization approach (Dunham, 2003).

Before building good predictive models one must understand the data. Summarization and involves methods for finding a compact description for a subset of data by gathering a variety of numerical summaries (including descriptive statistics such as averages, standard deviations and graphs) and looking at the distribution of the data (Thearling et.al ,2003).

Association Rules: - is a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. The investigation of relationships between items over a period of time is also often referred to as Sequence Analysis

Sequence Analysis: is used to determine sequential patterns in data .The patterns in the dataset are based on time sequence of actions, and they are similar to association data, however the relationship is based on time. In Market Basket analysis, the items are to be purchased at the same time, on the other hand, for Sequence Analysis the items are purchased over time in some order (Dunham, 2003).

2.1.7.2.1. Clustering techniques

Clustering is defined as the processes of creating a partition so that all the members of the partition are similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Clustering divides a database into different from each other, and whose members are very similar to each other (Alan R., 1995). Unlike classification, one doesn't know what the clusters will be when clustering start, or by which attributes of the data will be clustered, it is unsupervised learning.

Clustering is a technique useful for exploring data. It is particularly useful where there are many cases and no obvious natural groupings. Here, clustering data mining algorithms can be used to find whatever natural groupings may exist. Clustering analysis identifies clusters embedded in the data. A cluster is a collection of data objects that are similar in some sense to one another. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words, members of a cluster are more like each other than they are like members of a different cluster. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build predictive models.

Clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are formed. The term "class" is in fact frequently used as synonym to the term "cluster" (Dunham, 2003).

In this task the goal is to generate descriptions of classes of data (clusters), as opposed to classification, in which the classes are known beforehand. When identifying clusters, various Euclidean distance measures are used to compute how close are data items to each other in an n-dimensional space defined by the fields in data records(Fayyd et al, 1996) .

An entire collection of clusters is commonly referred to as a clustering, and there are various types of clustering's such as: hierarchical (nested) versus partitional (unnested), exclusive versus overlapping versus fuzzy, and complete versus partial (Han and Kamber, 2001).

Hierarchical versus Partitioned The most commonly discussed distinction among different types of clustering is whether the set of clusters is nested or un nested, or in more traditional terminology, hierarchical or partitioned. A partitioned clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. If we permit clusters to have sub clusters, then we obtain a hierarchical clustering, which is a set of nested clusters that are organized as a tree. Each node (cluster) in the tree (except for the leaf nodes) is the union of its children (sub clusters), and the root of the tree is the cluster containing all the objects. Often, but not always, the leaves of the tree are singleton clusters of individual data objects. Finally, note that a hierarchical clustering can be viewed as a sequence of partition clustering and a partitioned clustering can be obtained by taking any member of that sequence; i.e., by cutting the hierarchical tree at a particular level (Han and Kamber, 2001).

Exclusive versus Overlapping versus Fuzzy There are many situations in which a point could reasonably be placed in more than one cluster, and these situations are better addressed by non-exclusive clustering. In the most general sense, an overlapping or non-exclusive clustering is used to reflect the fact that an object can simultaneously belong to more than one group (class). For instance, a person at a university can be both an enrolled student and an employee of the university. A non-exclusive clustering is also often used when, for example, an object is “between” two or more clusters and could reasonably be assigned to any of these clusters.

Rather than make a somewhat arbitrary assignment of the object to a single cluster, it is placed in all of the “equally good” clusters. In a fuzzy clustering, every object belongs to every cluster with a membership weight that is between 0 (absolutely doesn’t belong) and 1 (absolutely belongs). In other words, clusters are treated as fuzzy sets. (Mathematically, fuzzy set is one in which an object belongs to any set with a weight that is between 0 and 1. In fuzzy clustering, we often impose the additional constraint that the sum of the weights for each object must equal 1.) Similarly, probabilistic clustering techniques compute the probability with which each point belongs to each cluster, and these probabilities must also sum to 1. Because the membership weights or probabilities for any object sum to 1, a fuzzy or probabilistic clustering does not address true multiclass situations, such as the case of a student employee, where an object belongs to multiple classes. Instead, these approaches are most appropriate for avoiding the arbitrariness of assigning an object to only one cluster when it may be close to several. In practice, a fuzzy or probabilistic clustering is often converted to an exclusive clustering by assigning each object to the cluster in which its membership weight or probability is highest.

Complete versus Partial a complete clustering assigns every object to a cluster, whereas a partial clustering does not. The motivation for a partial clustering is that some objects in a data set may not belong to well-defined groups. Many times objects in the data set may represent noise, outliers, or “uninteresting background.” For example, some newspaper stories may share a common theme, such as global warming, while other stories are more generic or one-of-a-kind. Thus, to find the important topics in last month’s stories, we may want to search only for clusters of documents that are tightly related by a common theme. In other cases, a complete clustering of the objects is desired. For example, an application that uses clustering to organize documents for browsing needs to guarantee that all documents can be browsed (Han and Kamber, 2001).

2.1.7.2.1.1. K-Means

The K-means algorithm, probably the first one of the clustering algorithms proposed, is based on a very simple idea: Given a set of initial clusters, assign each point to one of them, and then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point (Nanda, 2010).

In k-Means, the cancroids are computed as the arithmetic mean of the cluster all points of a cluster. The distances are computed according to a given distance measure, e.g. Euclidean distance. Although K-means has the great advantage of being easy to implement, it has two big drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success (Nanda, 2010).

2.1.7.3. Predictive vs. Descriptive models

A Predictive model makes a prediction about values of data using known results found from different data while the Descriptive model identifies patterns or relationships in data. Unlike the

predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties.
Clustering models are different from predictive models in that the outcome of the process is not guided by a known result, that is, there is no target attribute. Predictive models predict values for a target attribute, and an error rate between the target and predicted values can be calculated to guide model building. Clustering models, on the other hand, uncover natural groupings (clusters) in the data. The model can then be used to assign groupings labels (cluster IDs) to data points (Han and Kamber, 2001).

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions. Predictive modeling determines the goals for descriptive modeling, and the results of descriptive modeling guide the predictive modeling.

2.1.8. Data Mining Models application areas

In practical sense, the KDD generalizes the application of the process to non-database sources, although it emphasizes them as a primary source of data and also concerns the support for learning and analyzing the application domain (Lori, 2006).

KDD applications in the real world can be as diverse as the real world databases that exist today. The requirements for the KDD process in the variety of application domains, and even within the one domain of application, vary greatly. In insurance applications, for instance, discovery of rules for premium setting and fraud detection require quite different KDD processes even though the data is often very similar. Such different requirements entail a variety of techniques to be employed. In current practice the diversity of databases, application domains, and requirements has resulted in an artificial barrier to the promotion of KDD technology—trial-and-error has become a dominant approach with little, but increasing, formal guidance available. A theory or methodology will guide the KDD practitioner in the effective use of the technology (Matheus, et, all, 1993). KDD is widely used data mining technique that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results and provides popular application like marketing basket analysis, fraud detection, model visualization, exploratory data analysis telecommunication, and manufacturing (Woldekidan, 2003).

As Fayyd et al, (1996) discuss , KDD techniques can be applied in many domains, specifically on business information: marketing and sales data analysis, investment analysis, loan approval, fraud detection, manufacturing information, controlling and scheduling, network management, experiment result analysis, scientific information: sky survey cataloguing, biosequence Databases, geosciences: personal information within the organization.

Data mining is an interdisciplinary field with wide and diverse applications. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities: automated prediction of trends, behaviors and discovery of previously unknown patterns. Generally DM provides a vital role for different disciplines, few of them are:

- ✓ Science: Astronomy, bioinformatics, medicine discovery
- ✓ Business: Advertising, customer relationship management, investment, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care
- ✓ Web: Search engines
- ✓ Government: Law enforcement, profiling tax cheaters, anti-terror

Specifically it also provides different application in the organization as: Financial data analysis(loan payment prediction , classification and clustering of customers for targeted marketing, detection of money laundering and other financial crimes), retail industry(identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service), telecommunication industry(understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, improve the quality of service), biological data analysis(Gene and DNA sequences, semantic integration of heterogeneous, distributed genome databases), etc.

2.1.9. A Comparative Study of Data Mining Models

SEMMA and CRISP-DM are usually referred as methodologies or processes in many literatures, in the sense that they consist of a particular course of action intended to achieve a result. According to Matheus, et, al (1993), KDD and SEMMA process models are compared with respect to their stages and tasks, almost they are equivalent: Sample can be identified with Selection; Explore can be identified with Preprocessing; Modify can be identified with Transformation; Model can be identified with data mining; Assess can be identified with Interpretation/Evaluation. Examining it thoroughly, it can be affirmed that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process.

Whereas when we comparing the KDD stages with the CRISP-DM stages is not as straightforward as in the SEMMA situation. Nevertheless, we can first observe that the CRISP-DM methodology incorporates the steps that must precede and follow the KDD process. But we can compare the KDD and CRISP-DM models by their phases as follows.

The Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user; The Deployment phase can be indentified with the consolidation by incorporating this knowledge into the system.

Concerning the remaining stages, we can say that: The Data Understanding phase can be identified as the combination of Selection and Preprocessing: The Data Preparation phase can be identified with Transformation; the Modeling phase can be identified with DM; the Evaluation phase can be identified with Interpretation/Evaluation (Nanda, 2010).

2.2. Related Works

In our country, there are works done to assess the application of DM in the different sectors like Airlines, Banking, Insurance, HealthCare, and Customs. Henock (2002) and Denekew (2003) for example, conducted a research on the application of DM for customer relationship management in the airlines industry as a case study on Ethiopian Airlines. Both Henock and Denekew used clustering and classification techniques with k-Means and decision tree algorithms. In addition, Kumneger (2006) has also tried to study the application of DM techniques to support customer relationship management for the Ethiopian Shipping Lines. Kumneger has applied clustering and classification techniques with k-Means and decision tree algorithms.

In addition, Leul (2003) tried to apply the DM techniques for crime prevention as a case study on the Oromia Police Commission. Leul used the classification technique, decision tree and neural network algorithms to develop the model, which will help to classify crime records.

To the knowledge of the researcher there are only three attempts in our country that have been done so far towards the application of DM in the children related. Shegaw (2002) conducted a research on the application of DM in predicting child mortality in Ethiopia as a case study in the Butajira Rural Health Project. Shegaw employed the classification technique, neural network and decision tree algorithms to develop the model for predicting child mortality. The second research done by Helen (2003) also tried to study the application of DM technology to identify significant patterns in census or survey data as a case of the 2001 child labor survey in Ethiopia. She has applied the association rule DM technique and the Apriori algorithm for identifying relationships between attributes within the 2001 child labor survey database that she used to clearly understand the nature of child labor problem in Ethiopia. Apart from the association rule technique the expectation maximization-clustering algorithm were used to categorize the final selected datasets. The other research, which is undertaken by Woldekidan (2003), tried to apply the DM techniques for street children Ethiopia. He used the Apriori algorithm for Association rule mining.

These researchers Shegaw ,Helen and Woldekidan have tried to apply the different DM techniques in support of children classification. But as per the knowledge of the researcher till now there is no any work that have been done so far regarding the application of DM in orphan and vulnerable children in the children related organizations in Ethiopia. Hence this study has a great contribution in applying DM technology for the purpose of children classification to different categories in children support organization to get funds from different donors to help those vulnerable children's .

2.3. Orphan and Vulnerable Children in Ethiopia

Many people are aware that AIDS kills, but many are also ignorant of the result that AIDS deaths have on the families left behind. As children are left orphaned, forced to take responsibility for siblings, lead households, and provide for those remaining, they become extremely vulnerable. Psychological trauma deeply affects orphaned children, who may suffer from stress, depression, and a great deal of hopelessness. Limited or constrained resources may prevent these children from attending school and make them extremely susceptible to exploitation and abuse at the hands of those who appear to be well meaning (National Children's Commission, 1988).

One of the greatest challenges to child-led households is poverty. To address this issue, Love for Children Organization offering seed funds to communities and training, where it is wanted, in order to help foster income-generating activities. In other words, sustainability is the name of the game, and in Ethiopia, these programs have been rewarding and successful (National Children's Commission, 1988).

Education remains inaccessible in many parts of Sub-Saharan Africa to child-headed families that cannot afford the costs of supplies or uniforms. Therefore, advocating for universal free education is a priority, as school settings provide a safe haven for children where social stigmas may be left at the door (Global Action for Children, 2005).

It is heart-warming to hear that Love for Children Organization has seen successes in many of their initiatives with orphans and vulnerable children, from the implementation of non-discriminatory by-laws towards people living with AIDs in Ethiopia to the development of HIV/AIDS education programs in primary and secondary schools (UNICEF and USAID, 2007).

2.3.1. Concepts Related to OVC

The concept OVC is defined by UNICEF as one who has lost one or both parents being referred to as an orphan and a vulnerable child as one whose parents are terminally ill UNICEF (2006) The concepts of OVC will be reviewed so as to show guideline about OVC.

This could be said with the children that are found in Love for Children Organization compound as those who have lost either one or both parents. Vulnerability is stated according to the poverty levels in Love for Children Organization compound and what the community thinks it should be.

2.3.2. Who is a Child?

The United Nations Convention on the Rights of a child defines a child as any person below 18 years of age. According to UNICEF an orphan is one who has lost one or both parents UNICEF (2006) defines orphans and vulnerable children as "children who are compromised as a result of the illness or death of an adult who contributed to the care and/or financial support of the child".

2.3.3. Definition of OVC

According to Global Action for Children an orphan is a child who is below 18 years and who has lost one or both parents (Global Action for Children, 2005). The World Bank categorizes OVC in the context of war affected children, orphans, street children, AIDS affected children, disabled children and child labourers (Children's Rights Analysis, 2006). Furthermore Skinner et al (2004) writes that the term 'orphaned and vulnerable children' was introduced due to limited usefulness of the right definition of orphan hood in the scenario of HIV /AIDS.

The working paper of (UNICEF and USAID, 2007) reviewed the status of orphans and categorized them as "children who are without parental guardianship or care". A maternal orphan is one who has lost a mother and a paternal orphan is one who has lost a father. A double orphan is one who has lost both parents. When comparing with the current trend in Love for Children Organization compound it should be stated that these definitions do apply as vulnerability does vary according to the type of orphan hood one has been classified in.

2.3.4. Factors Contributing to OVC

Different factors continue to contribute to orphans and vulnerable children, Therefore some of the factors are listed below;

2.3.4.1. Diseases and Poverty

Poverty is state of not having the necessary basic needs to maintain human life and human health. It means scarcities and deficiencies. A poverty line is a tool to measure poverty and for separating the poor from the non-poor. A poverty line is constructed according to the value of income or consumption necessary to maintain a minimum standard of human nutrition and other basic necessities (Feuerstein, 1997). Most children who are dropping out of school are under the poverty line and failing to meet their basic needs such as clothes, food and other necessities to continue their life and schooling.

It has been stated by Mwaanga (unpublished thesis, 2003) that the majority of citizens living in low-income areas possibly will be living below the poverty line. This is the population which is subjected to whom multiple deprivations, such as lack of food, shelter, health care, education and employment. They undergo problems from lack of assets, items and opportunities which money can provide.

For such people, the challenges of every day survival are never-ending and achieving a reasonable standard of life and health is difficult. The load of such a life is more often than not laid upon mothers and children especially orphans and vulnerable children (Feuerstein, 1997).

Furthermore HIV and AIDS is one of the pandemics in Ethiopia that has attributed to the vicious poverty cycle where poor parents who die young, leaving behind a trail of orphans, pass on their poverty to their children. It has been stated that with the current world order (market economy) and the AIDS epidemic more people are losing their fragile grip and falling below the poverty line. It is estimated that 1.3 billion people – more than a fifth of all people in the world –live in poverty (Feuerstein, 1997).

2.3.4.2. Nutrition

Most of the time that lack of food and basic meals was a contributing factor in OVC dropping out of school. According to Global Action for Children, malnutrition is one of the elevated risks for orphans and vulnerable children (Global Action for Children, 2005). Furthermore Irish Aid

(2008, p.9) report states that malnutrition and poor health is a large contributor to low retention and poor performance in school.

2.4. Data Mining Methods for Orphan and Vulnerable Children's

Data mining functionalities are used to specify the kind of pattern to be found in data mining tasks (Alan R. 1995). Data mining methods or techniques may be classified by the function they perform or according to the class of application they can be used in. DM tasks depend on the kind of knowledge that the KDD/DM system looks for. Each DM task has its specific features and follows specific steps in the discovery process (Nanda, 2010).

A database is a store of information but more important is the information which can be inferred from it. In order to do this a wide variety of data-mining methods or techniques should be used. There is no particular rule that would tell you when to choose a particular technique over another one. Sometimes those decisions are made relatively arbitrary based on the availability of data mining analysts who are most experienced in one technique over another. These techniques can be used for either discovering new information within large databases or for building predictive models (Thearling et.al, 2003). The following are some of data mining techniques that are used in most cases.

2.4.1. Classification techniques

The task is to discover whether an item from the database belongs to one of some previously defined classes. The main problem, however, is how to define classes. In practice, classes are often defined using specific values of certain fields in the data records or some derivatives of such values (Fayyd et al, 1996).

Classification is learning a function that maps (classifies) a data item into one of several predefined classes. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from a historical database. The main objective of classification is to identify the characteristic that indicate the group to which each case belong. This pattern can be used both to understand the existing data and to predict how new instance will behave. That is the system take a case or records with certain known attribute values and able to predict what class this case belongs to. Prediction can be viewed as the construction and use of a model to asses the class of unlabeled sample is likely to have (Witten and Frank, 2005). That means, it predict unknown or missing class value.

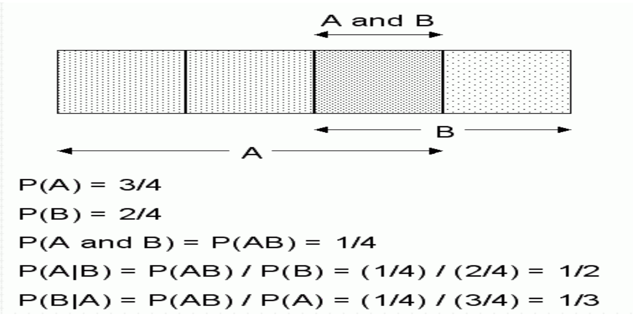
2.4.1.1. Naive Bayes Algorithm

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data (Nanda, 2010).

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

Bayes' Theorem: Prob(B given A) = Prob(A and B)/Prob(A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone.



$P(A) = 3/4$

$P(B) = 2/4$

$P(A \text{ and } B) = P(AB) = 1/4$

$P(A|B) = P(AB) / P(B) = (1/4) / (2/4) = 1/2$

$P(B|A) = P(AB) / P(A) = (1/4) / (3/4) = 1/3$

Figure 2.3 a Scenario that shows how Bayesian network is constructed

Naive Bayes makes the assumption that each predictor is conditionally independent of the others. For a given target value, the distribution of each predictor is independent of the other predictors.

In practice, this assumption of independence, even when violated, does not degrade the model's predictive accuracy significantly, and makes the difference between a fast, computationally feasible algorithm and an intractable one.

Sometimes the distribution of a given predictor is clearly not representative of the larger population when training the model.

The Naive Bayes algorithm affords fast, highly scalable model building and scoring. Naive Bayes can be used for both binary and multiclass classification problems (Nanda, 2010).

2.4.1.2. Decision Trees

A Decision Tree is a tree-structured plan of a set of attributes to test in order to predict the output. To decide which attribute should be tested first, simply find the one with the highest information gain (Nanda, 2010).

Decision trees extract predictive information in the form of human-understandable rules. The rules are if-then-else expressions; they explain the decisions that lead to the prediction.

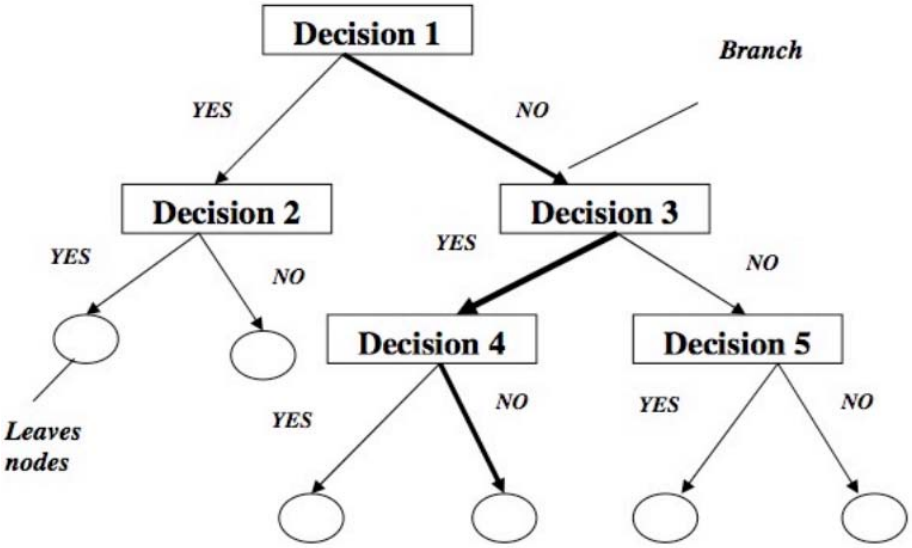


Figure 2.4.A Scenario that shows how decision tree is constructed

Decision trees use simple knowledge representation to classify examples into a finite number of classes. In a typical setting, the tree nodes represent the attributes, the edges represent the possible values for a particular attribute, and the leaves are assigned with class labels (Barbara et al., 2001)

Decision trees defined as simple knowledge representation and they classify examples/records to a finite number of classes, the nodes are labeled with attribute names, the edges are labeled with possible values for this attribute and the leaves labeled with different classes (Witten and Frank, 2005).Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes.

Because of their tree structure and ability to easily generate rules decision trees are the favored technique for building understandable models. Decision trees are also a natural choice when the goal is to generate rules that can be easily understood, explained, and translated into SQL or natural language (Witten and Frank, 2005).

Decision trees are a wonderfully versatile tool for data mining. There are two main types of decision trees (Thearling et.al, 2003):

1. Classification trees- Takes categorical values and label records and assign them to the proper class. The classification tree reports the class probability, which is the confidence that a records is in a given class.
2. Regression tree- Estimates the value of a target variable that takes on numeric values.

All decision tree construction methods are based on the principle of recursively partitioning the dataset until homogeneity is achieved. Various listed decision tree algorithms such as CHAID (Chi-Square Automatic Interaction Detector), C4.5/C5.0, CART (Classification and Regression Trees), ID3 and many others with less familiar algorithms produce trees that differ from one another in the number of splits allowed at each level of tree, how those splits are chosen when the tree is built (Witten and Frank, 2005).

2.4.1.3. Neural Networks

As discussed by (Witten and Frank, 2005) neural networks have the topology of a directed graph and loosely simulate the structure of biological neural networks in human brains. They are composed of processing nodes that convey actions to each other using connections

Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e. the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs that simply follow instructions in a fixed sequential order.

Neural networks are good choice for most classification and prediction tasks when the results of the model are more important than understanding how the model works. Neural networks actually represent complex mathematical equation, which lots of summation, exponential functions, and many parameters. These equations describe the neural network, but are quite opaque to human eyes. The equation is the rule of the network, and it is useless for our understanding.

Neural networks can produce very good predictions i.e. the advantages of neural network is their highly accurate predictive models that can be applied across a large number of different types of problems. But they are neither easy to use nor easy to understand, that is the results are difficult to understand because a neural network is a complex nonlinear model that doesn't produce rules (Khalid S. 2003).

The broad applicability of neural networks to real world business problems have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs including sales forecasting, industrial process control, customer research, data validation, risk management etc (Witten and Frank, 2005).

Adaptive Bayes Network (ABN) is an Oracle proprietary algorithm that provides a fast, scalable, non-parametric means of extracting predictive information from data with respect to a target attribute (Nanda, 2010).

ABN, in single feature build mode, can describe the model in the form of human-understandable rules. The rules produced by ABN are one of its main advantages over Naive Bayes. The business user, marketing professional, or business analyst can understand the basis of the model's predictions and can therefore be comfortable acting on them and explaining them to others. In addition to explanatory rules, ABN provides performance and scalability, which are derived via a collection of user parameters controlling the trade-off of accuracy and build time ABN predicts binary as well as multiclass targets. Binary targets are those that take on only two values, for example, *buy* and *not buy*. Multiclass targets have more than two values, for example, products purchased (product A or product B or product C) (Nanda, 2010).

Chapter Three

Data Preparation

3.1. Overview

In this research, the supervised (decision tree , naive bayes and neural network) DM tasks are experimented and this chapter discusses about the data preparation or pre processing tasks of the data mining in addition to solve the problems mentioned above (in chapter one), the following approaches have been implemented. Those are embraces detail description of the database from which the research data has been collected. The data cleaning, transformation and integration activities are explained in this chapter.

3.2. Data Understanding

Data mining requires collecting great amount of data (available in warehouses or databases) to achieve the intended objective.

The initial dataset is collected from the Love for children organization. The data set selected comprises child data the organization has been using from 2000 to 2011. The data has been imported from the Microsoft Access database of the organization to Microsoft Excel for processing. Several preprocessing methods have been applied to get the relevant data for the system. The data has more than 20 attributes like (Child _Name', 'House _Number', 'Telephone _Number', Sex, Age, Status _of _ Natural _Parents, Died _Parent, Guardian _Occupation, Cause _of _Parent _Death, Guardian _Relation _with _child, Child_ Duties _at_ Home, Child _Health_ Condition, Child _Grade_ Level, Child _Type) and are total of more than 20000 records. The data set includes not only the children who have got fund but also whose application has been rejected.

The data which are used for analyzing the specified case are not complete. Some data has been removed from the database because it is not that much significant in the final result. In addition, attribute name changes and data reshuffling had performed to have the required dataset. A total of 11 attributes and 17044 records have taken for analysis. This dataset is sufficient enough to perform some data mining techniques to get the desired result.

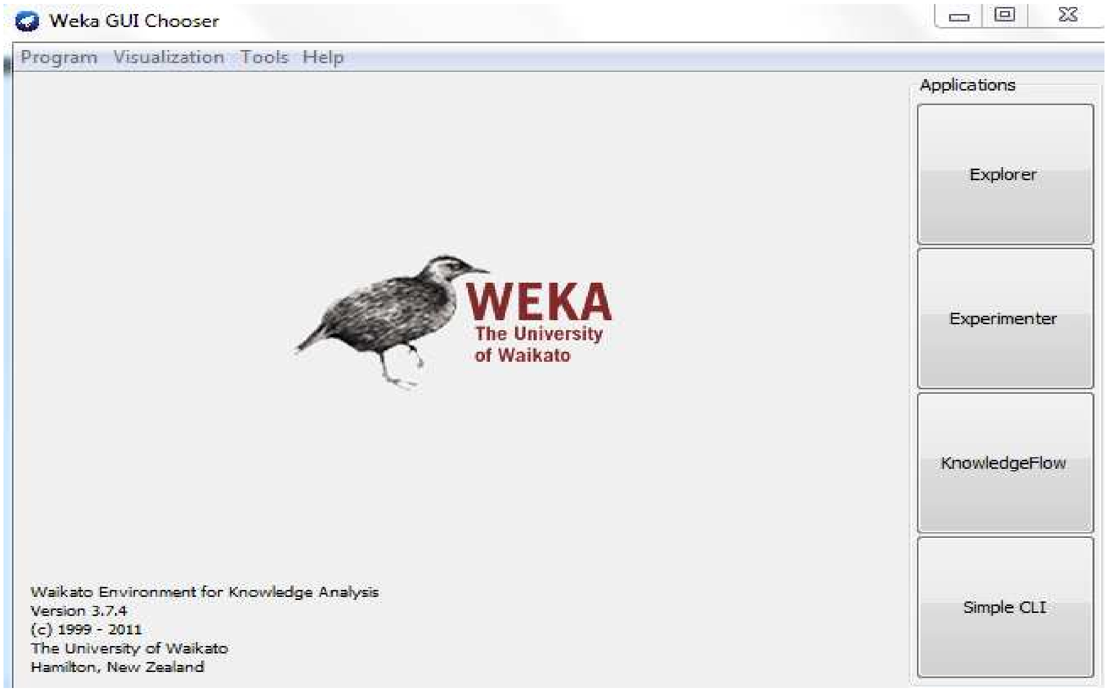
The selected attributes are the target attributes which are used for classification purpose. These are: Sex, Age, Status _of _ Natural _Parents, Died _Parent, Guardian _Occupation, Cause _of _Parent _Death, Guardian _Relation _with _child, Child_ Duties _at_ Home, Child _Health_ Condition, Child _Grade_ Level, Child _Type.

3.3. Implementation Tool

The Waikato Environment for Knowledge Analysis (Weka) was utilized to perform the mining of data using the classification techniques . Weka is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and Weka has been tested under Linux, Windows, and Macintosh operating systems. Java provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset (Rogers, 2001).

Weka consist collection of machine learning algorithms for solving real-world data mining problems. The package has three different interfaces: a command line interface, an Explorer GUI interface (which allows one to try out different preparation, transformation and modeling algorithms on a dataset), and an Experimenter GUI interface (which allows to run different algorithms in batch and to compare the results) (Witten & Frank, 2005). As one of the functionalities of the Weka software, classification and the decision tree, Bayesian network and neural network algorithm in particular is supported (Witten & Frank, 2005).

The obvious advantage of a package Weka is that a whole range of data preparation, feature selection and data mining algorithms are integrated. This means that only one data format is needed, and trying out and comparing different approaches becomes really easy. The package also comes with a GUI, which should make it easier to use (Witten & Frank, 2000). The following figure depicts the graphical user interface chooser.



Waikato Environment for Knowledge Analysis
Version 3.7.4
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Figure 3.1: Weka GUI Chooser

3.4. The Knowledge Discovery Process

According to Fayyd et al, the KDD can be viewed as constituting four major phases/activities. These are data selection, data preprocessing, data mining, and post-processing/interpretation.

3.4.1. Data Selection

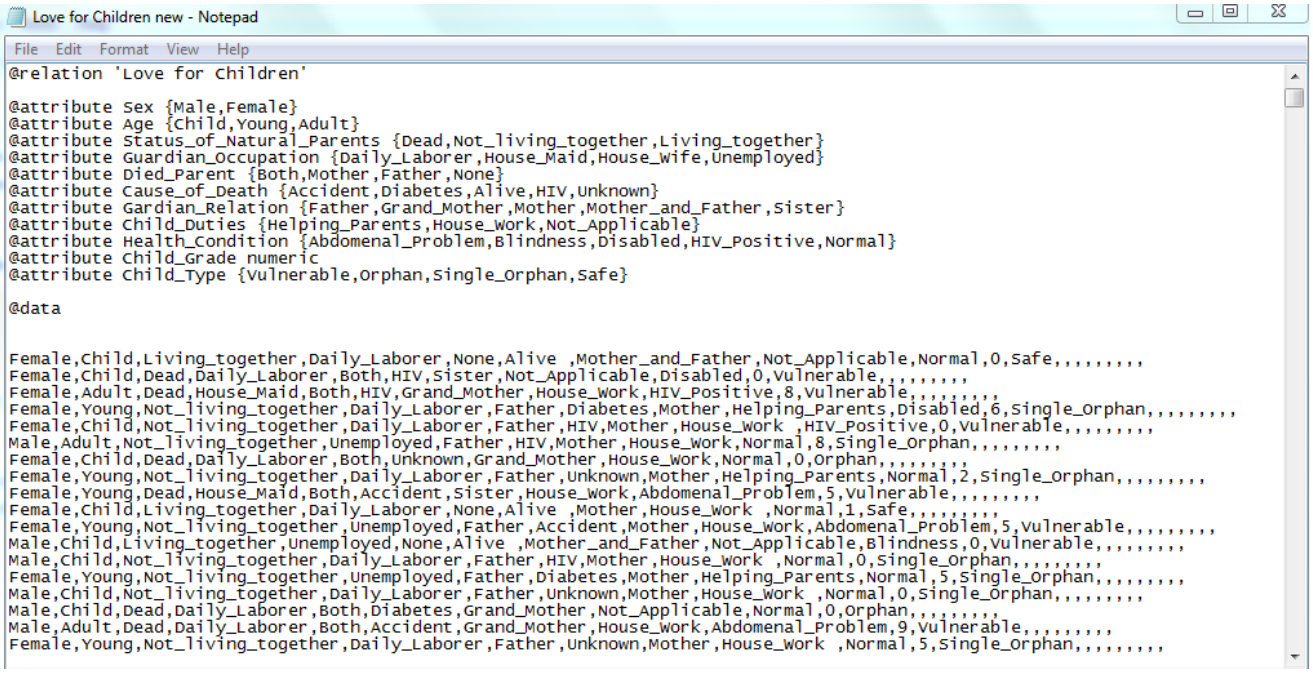
Data selection is important as the data consists of features that are not related to the problem at hand such as children phone_ number, house_ number and child _name since some of them are redundant, and irrelevant, or invariable throughout the dataset. Taking into account such features in the automated analysis might not result in meaningful patterns. Hence, 11 attributes out of the original 20 that best suit the objectives of the research were selected

3.4.2. Data Processing

The data processing part is one of the challenging parts of the research. Its preparation has taken much of the research time and effort as described by Han and Kamber (2001) the data preprocessing task of data mining includes data cleaning, data integration, data transformation and data reduction. Each of these tasks has been performed by the researcher in the source of preparing the data for model building.

3.4.2.1. Dataset Format

The WEKA system uses a common file format to store its datasets and thus presents the user with a consistent view of the data regardless of what machine learning scheme may be used. This file format, the Attribute-Relation File Format (ARFF), defines a dataset in terms of a relation or table made up of attributes or columns of data. Information about the names of the relation, and the data types of the attributes are stored in the ARFF header, with the examples or instances of data being represented as rows of data in the body of the ARFF file (Garner, nd). Since the data mining software used to generate classification accepts data only in arff format, the researcher first converted the data on Ms Excel file into comma separated text format (.csv) format which is a format where commas are placed between values in adjacent columns. The database was then opened in Word, header information added. That is, the @ symbol was placed in front of the relation name and each attribute. The @DATA symbol was placed before the data. Finally, saving the file extension was changed to .arff format. Data in arff format is then given to Weka software. The following figure indicates the representation of the data in ARFF format.



```
@relation 'Love for Children'

@attribute Sex {Male,Female}
@attribute Age {Child,Young,Adult}
@attribute Status_of_Natural_Parents {Dead,Not_living_together,Living_together}
@attribute Guardian_Occupation {Daily_Laborer,House_Maid,House_Wife,Unemployed}
@attribute Died_Parent {Both,Mother,Father,None}
@attribute Cause_of_Death {Accident,Diabetes,Alive,HIV,Unknown}
@attribute Gardian_Relation {Father,Grand_Mother,Mother,Mother_and_Father,Sister}
@attribute Child_Duties {Helping_Parents,House_Work,Not_Applicable}
@attribute Health_Condition {Abdomenal_Problem,Blindness,Disabled,HIV_Positive,Normal}
@attribute Child_Grade numeric
@attribute Child_Type {Vulnerable,Orphan,Single_Orphan,Safe}

@data


Female,Child,Living_together,Daily_Laborer,None,Alive ,Mother_and_Father,Not_Applicable,Normal,0,Safe,,,,,,,,
Female,Child,Dead,Daily_Laborer,Both,HIV,Sister,Not_Applicable,Disabled,0,Vulnerable,,,,,,,,
Female,Adult,Dead,House_Maid,Both,HIV,Grand_Mother,House_Work,HIV_Positive,8,Vulnerable,,,,,,,,
Female,Young,Not_living_together,Daily_Laborer,Father,Diabetes,Mother,Helping_Parents,Disabled,6,Single_Orphan,,,,,,,,
Female,Child,Not_living_together,Daily_Laborer,Father,HIV,Mother,House_Work ,HIV_Positive,0,Vulnerable,,,,,,,,
Male,Adult,Not_living_together,Unemployed,Father,HIV,Mother,House_Work,Normal,8,Single_Orphan,,,,,,,,
Female,Child,Dead,Daily_Laborer,Both,Unknown,Grand_Mother,House_Work,Normal,0,Orphan,,,,,,,,
Female,Young,Not_living_together,Daily_Laborer,Father,Unknown,Mother,Helping_Parents,Normal,2,Single_Orphan,,,,,,,,
Female,Young,Dead,House_Maid,Both,Accident,Sister,House_Work,Abdomenal_Problem,5,Vulnerable,,,,,,,,
Female,Child,Living_together,Daily_Laborer,None,Alive ,Mother,House_Work ,Normal,1,Safe,,,,,,,,
Female,Young,Not_living_together,Unemployed,Father,Accident,Mother,House_Work,Abdomenal_Problem,5,Vulnerable,,,,,,,,
Male,Child,Living_together,Unemployed,None,Alive ,Mother_and_Father,Not_Applicable,Blindness,0,Vulnerable,,,,,,,,
Male,Child,Not_living_together,Daily_Laborer,Father,HIV,Mother,House_Work ,Normal,0,Single_Orphan,,,,,,,,
Female,Young,Not_living_together,Unemployed,Father,Diabetes,Mother,Helping_Parents,Normal,5,Single_Orphan,,,,,,,,
Male,Child,Not_living_together,Daily_Laborer,Father,Unknown,Mother,House_Work ,Normal,0,Single_Orphan,,,,,,,,
Male,Child,Dead,Daily_Laborer,Both,Diabetes,Grand_Mother,Not_Applicable,Normal,0,Orphan,,,,,,,,
Male,Adult,Dead,Daily_Laborer,Both,Accident,Grand_Mother,House_Work,Abdomenal_Problem,9,Vulnerable,,,,,,,,
Female,Young,Not_living_together,Daily_Laborer,Father,Unknown,Mother,House_Work ,Normal,5,Single_Orphan,,,,,,,,
```

Figure 3.2: Representation of the data in ARFF format

3.4.2.2. Data Cleaning

To filter bad data, data purification is important. There was incomplete record and unnecessary attributes for the system when this data was collected. Removing and completing these records was the main task. The data mining algorithms perform quality mining as long as there is a cleaned data, which is free from outliers, noise and no missing values for an attribute(s). Many of the attributes used in this research are derived from the base attributes.

However the researcher tried to do the data cleaning such as handling missing attributes values and noise removal before deriving the attributes that are used to build the classification model. Accordingly, there is one attribute from the selected that had missing values. The following pointes briefly describe the methods employed in handling missing and outlier data. The researcher handled these missing values by replacing with the user-defined value ‘None’. The data has been preprocessed so that it will be ready for mining.

3.4.2.3. Data Integration and Data Transformation

This process is integration of data from multiple databases, data warehouses, or files. The data format in Ms-Access has changed to Ms-excel for the use of weka knowledge discovery tool. Data mining often requires data integration (merging of data) from multiple data stores. May also need to be transformed in to form appropriate for mining (Han and Kamber, 2001).

The data transformation task includes smoothing, aggregation, generalization normalization and attribute selection (Han and Kamber, 2001). The data set hat has been used for modeling building are smoothened and normalized. In this paper, normalization of data has been processed.

3.4.2.4. Data Reduction Methods

There were more than 20000 instances and 20 attributes from the database. But as pre the need of the system, non relevant attributes have been removed and reduced into 11 attributes.

In all the data preprocessing tasks Ms-Access, Ms-Excel and Weka soft ware have been intensively employed and used. Since the imported data had missing values, it needs preprocessing. For example, the attribute 'Died_Parent' was not complete for those children whose both parents are alive. In this case, this are handled these missing values by replacing with the user-defined value 'None'.

The data has been preprocessed so that it will be ready for mining. Attribute reduction has also been done by eliminating some attributes which are not relevant for my project. For example, the researchers have been eliminated attributes like 'Child_Name', 'House_Number', 'Telephone_Number'.

3.5. Building Data Mining Model

WEKA version 3.7.4 has been used for building the model and visualizing the result. 66% of the data has been used for training and the remaining of the data has been used for testing because this is the preferred proportion that has been used on many literatures and as it is one of the options on WEKA.

As has been mentioned above, one of the core problems of the organization pertaining to child data classification is lack of appropriate mechanism to set sound criteria. To alleviate this problem, the researcher have tried to build a classification model using Naïve Bayes, Multilayer Perception Neural Network, and decision tree (C4.5) using WEKA. The detailed description of each of the classifiers is given below.

3.5.1. Naïve Bayesian Classifier

This classifier works on the Bayesian probability distribution of the data. The classifier evaluates absolute probabilities of the classes looking at the training data distribution. This information (the evaluated probability) is used to classify test data.

In this research, the researcher have experimented classification using Naïve Bayesian Classifier. Given children data, this classifier classifies the data object as vulnerable, single _ orphan, orphan or safe. Thus, by using the classification model and depending on the result of the classifier, the organization can decide whether to provide or deny fund to a particular child that applies for getting fund from the organization.

3.5.2. Neural Networks:

A multilayer perception is a feed-forward neural network with one or more hidden layers. The network consists of:

- An input layer of source neurons which introduces input values into the network.
- At least one middle or hidden layer of computational neurons that performs classification of features.
- An output layer of computational neurons that pass the output of the network on to the world outside the neural network.
- The input signals are propagated in a forward direction on a layer-by-layer basis.

In this research, the researcher have experimented classification using Multilayer Preceptor Neural Network that classifies the data object as vulnerable, single _orphan, orphan or safe using the given child data. Thus, in a similar fashion as the Naïve Bayes model described above, by using this classification model and depending on the result of the classifier, the organization can decide whether to provide or deny fund to a particular child.

3.5.3. Decision Tree (C4.5) Classifier

Decision trees represent a supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The WEKA classifier package has its own version of C4.5 known as J48. The researcher used decision tree (C4.5/J48) as one approach for classifying the data since it has the following benefits:

- Decision trees are easy to understand
- Decision trees are easily converted to a set of production rules
- Decision trees can classify both categorical and numerical data
- There are no a priori assumptions about the nature of the data.

3.6. Methods

After getting familiar with the problem domain, the data used for the paper is obtained from Love for Children Organization found in Addis Ababa sub city Mexico. The office keeps records of Children on a centralized manner using MS-Access in Amharic language with more than 20 attributes and 20000 records. Unnecessary attributes and some of records are removed and only important attributes are selected for this paper.

Among data mining techniques, classification has been used for the interpretation of data that enables to identify the children's data to orphans, single _orphan, vulnerable and safe. Classification is similar to clustering in that it also partitions records into distinct segments called classes. But unlike clustering, classification analysis requires that the end-user know in advance of time how classes are defined. It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes (Meera et al., 2010). For this, weka 3.7.4 data mining tool have been used. The convenient algorithm selected from classification data mining techniques is Decision tree, Bayesian network and neural network algorithms.

Chapter Four

Result of the Experimentation

This chapter presents steps and procedures followed during experimentations. The main objective of this research is, discovering regularities for predicting children classification within Love for children organizations dataset. Having this purpose in mind, the model building phase in the DM process of this investigation is carried out by the supervised classification techniques are adopted and this technique is implemented using Weka DM tool. In order to perform the model building process of this study, 17044 dataset is used for the classification models.

4.1. Classification Model Building

The algorithm selected for classification purposes were j48 decision tree, naive bayes and neural network they can classify an instance in to already identified class. The researcher tested the algorithm with different algorithm and record numbers to improve the classification accuracy. Finally, compared and selected the best classification model from the three algorithms.

This study is developing the predictive model using the classification techniques then once the dataset is ready to be used, the next step is building the classification model using the selected DM tool. As it was discussed before, the Weka version 3.7.4 software is used for conducting this study.

For starting the classification modeling experiments, the decision tree (in particular the J48 algorithm), the naïve bayes and neural network methods are selected. The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification modes. Classification is analyzed to measure the accuracy of the classifiers in categorizing the children classification into specified classes. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications. The classification accuracy of each of these models is reported and their performance is compared in classifying new instances of records.

4.1.1. J48 Decision tree Model building

At each node it will be sent either left or right according to some test. Eventually, it will reach a leaf node and be given the label associated with that leaf. Generally, this research is more interested in generating rules that best predict the children's classification to orphan, vulnerable, half-orphans and safe of children's.

As described before, the J48 algorithm is used for building the decision tree model. J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Initially the classification model is built with the default parameter values of the J48 algorithm. The following table summarizes the default parameters with their values for the J48 decision tree algorithm.

Parameter	Description	Default Value
Confidence Factor	The confidence factor used for the pruning (smaller values incur more pruning)	0.25
minNumObj	The minimum number of instance per leaf	2
Unpruned	Whether pruning is performed	False

Table 4.1 some of the J48 algorithm parameters and their default values

By changing the different default parameter values of the J48 algorithm, the experimentations of the decision tree model-building phase are carried out.

4.1.1.1. Experiment 1

The first experimentation is performed with the default parameters. The default 10-fold cross validation test option is employed for training the classification model. Using these default parameters the classification model is developed with a J48 decision tree having 20 numbers of leaves and 27 tree size. Table 4.2 depicts the resulting confusion matrix of this model.

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single_Orphan	Safe		
Vulnerable	7159	0	66	98	7323	98.13%
Orphan	0	1983	0	0	1983	100%
Single_Orphan	41	0	4574	0	4615	98.48%
Safe	0	0	0	3123	3123	100%
Total	7256	1983	4584	3221	17044	98.7972%

Table 4.2 confusion matrix output of the J48 algorithm with default value

As shown in the confusion matrix above, the J48 learning algorithm scored an accuracy of 98.7972 percent, which indicates that, out of the total number of records supplied, 16839(98.79 %) records are classified correctly and 205(1.21%) are misclassified or incorrectly classified, Furthermore, the resulting confusion matrix of this experiment has shown that 100% of the records are correctly classified in the orphans and safe. This shows that the algorithm classified the entire orphan and safe children are classified in their respective class and out of the 7323 vulnerable children, who are described in vulnerable, 7186 (98.13 %) of them are classified correctly in their designated class, i.e. vulnerable, while only 39 (0.0053 %) of them are misclassified in single _orphans and 98(0.013%) of them are misclassified in safe. In addition to this out of the 4615 single _orphan children, 4545 (98.48 %) of them are classified correctly in their designated class, i.e. single _orphan, while only 70(0.015 %) of them are misclassified in Vulnerable. As described before, the size of the tree and the number of leaves produced from this training was 27 and 20 respectively. Therefore, to make ease the process of generating rule sets or to make it more understandable, the researcher attempted to modify the default values of the parameters so as to minimize the size of the tree and number of leaves. With this objective, the minNumObj (minimum number of instances in a leaf) parameter was tried with 25, 20, 15, 10 and 5. But the minNumObj set to these it doesn’t give a better tree size and accuracy compared with the other trials. Means their value has no that much difference compared with the first one. That means the complexity of the decision tree to generate rules is the same in both the experiments. So, since there is no a tangible difference in the tree size and number of leaves in this experiments the accuracy of the model is 98.79% so, this experiment with the default minNumObj parameter value is taken as the J48 decision tree model.

4.1.1.2. Experiment 2

This experiment is performed, by changing the default testing option (the 10-fold cross validation). In this learning scheme a percentage split is used to partition the dataset into training and testing data. The purpose of using this parameter was to assess the performance of the learning scheme by increasing the proportion of testing dataset if it could achieved a better classification accuracy than the first experimentation. First this experiment has run with the default value of the percentage split (66%). The result of this learning scheme is summarized and presented in Table 4.3.

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	2401	0	31	41	2473	97.088%
Orphan	0	711	0	0	711	100%
Single_Orphan	0	0	1569	0	1569	100%
Safe	0	0	0	1042	1942	100%
Total	2401	711	1600	1083	6695	98.7575%

Table 4.3 Confusion Matrix output of the J48 algorithm with the percentage – split set to 66%

Out of the 17044 total records 11249(66%) of the records are used for training purpose while 5794(44%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 5794 testing records 5723(98.7575 %) of them are correctly classified. Only 72 (1.2425 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 2473vulnerable records 2401

(97.088%) of them are correctly classified while only 31 (0.0125%) of the records are misclassified in single –orphan and 41(0.0165) of the records are misclassified in safe. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan, single_ orphans and safe of their class. This shows that the model correctly classified those children data's in their respective class.

4.1.1.3. Experiment 3

This experiment is performed, by changing the default testing option (the 66% for training 44% for testing). So the percentage split parameter set to 70, which is to mean 70% for training and 30% for testing, resulted with a better accuracy. The result of this learning scheme is summarized and presented in Table 4.4.

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	2123	0	28	34	2185	97.16
Orphan	0	630	0	0	630	100%
Single_Orphan	0	0	1383	0	1383	100%
Safe	0	0	0	915	915	100%
Total	2123	630	1411	949	5113	98.7874%

Table 4.4 confusion Matrix output of the J48 algorithm with the percentage – split set to 70%

In this experiment out of the 17044 total records 11931 (70%) of the records are used for training purpose while 5113(30%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 5113 testing records 5051(98.7874 %) of them are correctly classified. Only 62 (1.2126 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 2185 vulnerable records 2123 (97.16%) of them are correctly classified while only 28 (0.0128%) of the records are misclassified in single –orphan and 34(0.0155) of the records are misclassified in safe. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan, single_ orphans and safe of their class. This shows that the model correctly classified those children data's in their respective class.

4.1.1.4. Experiment 4

In the fourth experiment on this is by changing the 70/30% to the default 50/50% parameters

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	3603	0	0	54	3657	98.52%
Orphan	0	1039	0	0	1039	100%
Single_Orphan	46	0	2261	0	2307	98.00%
Safe	0	0	0	1519	1519	100%
Total	3649	1039	2261	1573	8522	98.8266%

Table 4.5 Confusion Matrix output of the J48 algorithm with the percentage – split set to 50%

In this experiment out of the 17044 total records 8522 (50%) of the records are used for training purpose while 8522 (50%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 8522 testing records 8422(98.83 %) of them are correctly classified. Only 100 (1.17 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 3657 vulnerable records 3603 (98.52%) of them are correctly classified while only 54 (1.48%) of the records are misclassified in safe and out of 2307 single _orphan children's 2261(98.00%) are correctly classified as single _orphan's the rest 46(2.03%) are misclassified as vulnerable . Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class.

In the previous four experiments when the testing data is increased the performance of the algorithm for predicting the newly coming instances is also diminished as well when compare the 10-fold cross validation, 70/30% and 66/44% scores an accuracy of 98.797%,98.787% and 98.757% respectively but the 4th experiment changes this flow that is while using 50% of the records for testing 50% for training this scores highest accuracy 98.83%. Though this experiment is conducted by varying the value of the training and the testing datasets, this shows that the experiment 50/50% (98.83%), 10-fold cross validation (98.80%), 70/30% (98.79%), 66/44 % (98.76%) conducted, is better experiment from highest to lowest scoring respectively.

Generally, from the four experiments conducted before, the model developed with the 50/50% parameter values of the J48 decision tree algorithm test option gives a better classification accuracy of predicting of the children classification in their respective class category. Therefore, among the different decision tree models built in the foregoing experimentations, the fourth model, with the 50/50% parameter values, has been chosen due to its better overall classification accuracy.

4.1.2. Naïve Bayes Model Building

After experimenting with J48 decision tree algorithm with many parameter values, the best model that has shown better overall classification accuracy has been chosen. And to compare the result of the J48 decision tree classification model different naïve bayes experiments have been carried out.

The same attributes that are used to build the decision tree models, are also used in this naïve bays modeling experiments. With all preprocessing takes in place, the experimentation proceeded with the different naïve bayes models by changing the default parameter values.

The 10-fold cross validation, which is set by default, the percentage split with 66/44%, 70/30 % and 50/50% for training and testing the model test options are employed. Naïve Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes (Han and Kamber 2006).

4.1.2.1. Experiment 1

The first experiment of the naïve bayes model building is performed using the Naïve Bayes Simple algorithm with the default10-fold cross validation test option. Table 4.6 shows the resulting confusion matrix of the model developed using the Naïve Bayes Simple algorithm with the default 10-fold cross validation test option.

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	7113	0	112	98	7323	98.13%
Orphan	0	1983	0	0	1983	100%
Single_Orphan	97	0	4518	0	4615	97.68%
Safe	0	0	0	3123	3123	100%
Total	7210	1983	4630	3221	17044	98.1988%
10-fold cross validation						

Table 4.6 confusion matrix output of the naïve bayes simple algorithm

The result from this experiment shows that out of the 17044 total records 16737(98.1988%) of them are correctly classified. And 307 (1.8012 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 7323vulnerable records 7113 (9713%) of them are correctly classified while 112(0.015%) of the records are misclassified in single –orphan and 98(0.0134%) of the records are misclassified in safe. And out of 4615 single_ orphan records 4518(97.89) of them are correctly classified and 97(0.021%) of them are incorrectly classified as vulnerable. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class. The model developed with Naïve Bayes Simple Algorithm is less in the accuracy of classifying new children dataset to the respected class, compared with the decision tree model that is developed before.

4.1.2.2. Experiment 2

The second experiment of the naïve bayes model building is performed using the Naïve Bayes Simple algorithm with the 66/44% training and testing percentage split test option. Though different experiments are conducted by changing the size of the training and testing datasets, the one with 66/44% training and testing dataset scored better classification accuracy and it is presented here. The result of this experiment is shown in Table 4.7

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	2400	0	32	41	2473	97.05
Orphan	0	711	0	0	711	100%
Single_Orphan	30	0	1539	0	1569	98.089
Safe	0	0	0	1042	1042	100%
Total	2430	711	1571	1083	5795	98.2226%
Percentage split(66/44% Training and testing)test option						

Table 4.7 Confusion matrix output of the naïve bayes simple algorithm

As can be seen from the confusion matrix that resulted from the model developed by the Naïve Bayes Simple Algorithm with the 66/44% percentage split, the model scored an accuracy of 98.22%. This shows that from the total 5795 test data, 5692 (98.22%) of the records are correctly classified, while 103 (1.777%) of them are misclassified. In addition to this the confusion matrix also shows that from the total 2473 vulnerable records 2400 (97.05%) of them are correctly classified, while 32 (0.01294%) of the records are misclassified in the single _orphan and 41(0.01657%) are misclassified in safe. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class in addition , the confusion matrix shows that out of the total 1569 single _orphan children 1539 (98.089%) of the records are correctly classified in the single orphan's, while 30 (0.0191%) of them are incorrectly classified in vulnerable . Compared with the second classification the first classification is better in classifying children's data correctly.

4.1.2.3. Experiment 3

The third experiment of the naïve bayes model building is performed using the Naïve Bayes Simple algorithm with the 70/30% training and testing percentage split test option. Though different experiments are conducted by changing the size of the training and testing datasets, the one with 66/44% training and testing dataset scored better classification accuracy and it is presented here. The result of this experiment is shown in Table 4.8

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	2123	0	28	34	2185	97.16%
Orphan	0	630	0	0	630	100%
Single_Orphan	27	0	1356	0	1383	98.05%
Safe	0	0	0	915	915	100%
Total	2150	630	1384	949	5113	98.2593%
Percentage split(70/30% Training and testing)test option						

Table 4.8 confusion matrix output of the naïve bayes simple algorithm

The result from this experiment shows that out of the 5113 total records 5024(98.26%) of them are correctly classified. And 89(1.74 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 2185 vulnerable records 2123 (97.16%) of them are correctly classified while 28(1.28%) of the records are misclassified in single –orphan and 34(1.56%) of the records are misclassified in safe. And out of 1383 single_ orphan records 1356(98.05) of them are correctly classified and 27(1.95%) of them are incorrectly classified as vulnerable. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class. The model developed with 3rd experiment of the Naïve Bayes Simple Algorithm scores higher accuracy of classifying new children dataset to the respected class, compared with the experiment that is developed before.

4.1.2.4. Experiment 4

The fourth experiment of the naïve bayes model building is performed using the Naïve Bayes Simple algorithm with the 50/50% training and testing percentage split test option. Though different experiments are conducted by changing the size of the training and testing datasets, the one with 50/50% training and testing dataset scored better classification accuracy and it is presented here. The result of this experiment is shown in Table 4.9

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	3560	0	43	54	3657	97.35%
Orphan	0	1039	0	0	1039	100%
Single_Orphan	46	0	2261	0	2307	98.00%
Safe	0	0	0	1519	1519	100%
Total	3606	1039	2304	1573	8379	98.322%
Percentage split(50/50% Training and testing)test option						

Table 4.9 Confusion matrix output of the naïve bayes simple algorithm

As we can see from the confusion matrix of the model developed with this proportion, out of the 8522 testing records 8379(98.32 %) of them are correctly classified. Only 143 (1.68 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 3657 vulnerable records 3560 (97.35%) of them are correctly classified while only 54 (1.48%) of the records are misclassified in safe and 54(1.50%) are misclassified as safe and out of 2307 single _orphan children's 2261(98.00%) are correctly classified as single _orphan's the rest 46(2.03%) are misclassified as vulnerable. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class. Generally, the model developed with 4th experiment of the Naïve Bayes Simple Algorithm scores higher accuracy of classifying new children dataset to the respected class, compared with the experiment that is developed before in naïve bayes algorithm.

4.1.3. Neural Network Classification Model Building

After experimenting with many decision tree models, the best model that has been better overall classification accuracy has been chosen. To build the neural network model that classifies the data into the given classes based on the given data, it worked on finding the appropriate number of iterations that would result a maximum accuracy. The same attributes that are used to build the decision tree and naïve bayes models, are also used in this neural network modeling experiments. With all preprocessing takes in place, the experimentation proceeded with the different neural network models by having the default parameter values. The 10-fold cross validation, which is set by default, and the percentage split with 66-44%, 70/30% and 50/50% for training and testing the model test options are employed.

4.1.3.1. Experiment 1

This experiment is performed, by using the default testing option (the 10- fold cross validation testing option). The result of this learning scheme is summarized and presented in Table 4.10

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	7229	0	40	54	7323	98.716%
Orphan	0	1983	0	0	1983	100%
Single_Orphan	52	0	4563	0	4615	98.87%
Safe	44	0	0	3079	3123	98.59%
Total	7325	1983	4603	3133	17044	98.8152%
10-fold cross validation						

Table 4.10 Confusion matrix output of the neural network algorithm

The result from this experiment shows that out of the 17044 total records 16854(98.81%) of them are correctly classified. And 190 (1.1848 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 7323vulnerable records 7229 (98.716%) of them are correctly classified while 40(0.00546%) of the records are misclassified in single –orphan and 54(0.00737%) of the records are misclassified in safe. And out of 4615 single_orphan records 4563(98.87) of them are correctly classified and 52(0.01126%) of them are incorrectly classified as vulnerable. In addition out of 3123 safe children 3079(98.59%)of them are correctly classified and 44(0.0140%) are misclassified in vulnerable. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan of their class. This shows that the model correctly classified those children data's in their respective class.

4.1.3.2. Experiment 2

This experiment is performed, by changing the default testing option (the 10 cross validation testing option). So the percentage split parameter set to 66/44%, which is to mean 66% for training and 44% for testing, the result of this learning scheme is summarized and presented in Table 4.11.

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	2432	0	0	42	2474	98.30%
Orphan	0	711	0	0	711	100%
Single_Orphan	30	0	1539	0	1569	98.088%
Safe	0	0	0	1042	1042	100%
Total	2462	711	1539	1084	5796	98.7748%
Percentage split(66/44% Training and testing)test option						

Table 4.11 confusion matrix output of the neural network algorithm

The result from this experiment shows that out of the 5796 total records 5724(98.77%) of them are correctly classified. And 71 (1.22 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 2472vulnerable records 2432 (98.30%) of them are correctly classified while 42(0.01697%) of the records are misclassified in safe and . And out of 1539 single _orphan, records 1539(98.088%) of them are correctly classified and 30(0.019%) of them are incorrectly classified as vulnerable. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class.

Both of the two neural net models explained above have generally shown very good classification accuracy. However, the first model built using the default parameters and 10-fold cross validation excels both in overall accuracy. Hence, it is chosen as best neural net model.

The neural nets models are considered as black boxes. This is due to the fact that it does not really explicitly show why a certain record is segmented /classified in to a certain class. Besides, it does not generate rules like decision trees. Thus, no rules are derived for the model.

4.1.3.3. Experiment 3

This experiment is performed, by changing the default testing option (the 66% for training 44% for testing). So the percentage split parameter set to 70, which is to mean 70% for training and 30% for testing, The result of this learning scheme is summarized and presented in Table 4.12.

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	2151	0	0	34	2185	98.44%
Orphan	0	630	0	0	630	100%
Single_Orphan	27	0	1356	0	1383	98.05%
Safe	0	0	0	915	915	100%
Total	2178	630	1356	949	5113	98.807%
Percentage split(70/30% Training and testing)test option						

Table 4.12 Confusion matrix output of the neural network algorithm

The result from this experiment shows that out of the 5113 total records 5052(98.81%) of them are correctly classified. And 61 (1.18 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 2185 vulnerable records 2151 (98.44%) of them are correctly classified while 34(1.56%) of the records are misclassified in safe. And out of 1383 single_orphan records 1356 (98.05) of them are correctly classified and 27(1.95%) of them are incorrectly classified as vulnerable. In addition, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class.

4.1.3.4. Experiment 4

The fourth experiment is performed, by changing the testing option (the 70/ 44%). So the percentage split parameter set to 50/50%, which is to mean 50% for training and 50% for testing, the result of this learning scheme is summarized and presented in Table 4.13

Actual	Predicted				Total	correctly classified (accuracy rate)
	Vulnerable	Orphan	Single _Orphan	Safe		
Vulnerable	3560	0	43	54	3657	97.35%
Orphan	0	1039	0	0	1039	100%
Single_Orphan	0	0	2307	0	2307	100%
Safe	0	0	0	1519	1519	100%
Total	3560	1039	2350	1573	8522	98.86%
Percentage split(50/50% Training and testing)test option						

Table 4.13 Confusion matrix output of the neural network algorithm

As we can see from the confusion matrix of the model developed with this proportion, out of the 8522 testing records 8425(98.86 %) of them are correctly classified. Only 97 (1.14 %) records are incorrectly classified. In addition to this the resulting confusion matrix has shown that out of 3657 vulnerable records 3560 (97.35%) of them are correctly classified while only 43 (1.18%) of the records are misclassified in single _orphan and 54(1.48%) are incorrectly classified in safe. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the single _orphan, orphan and safe of their class. This shows that the model correctly classified those children data's in their respective class. In comparison of the neural network model experiments the fourth experiment that splitting to 50/50% scores highest.

The neural network model does not really explicitly show why a certain records are segmented/ classified in to a certain class. Besides it does not generate rules like decision trees.

4.2. Comparison of J48 Decision Tree and Naïve Bayes Models

Selecting a better classification technique for building a model, which performs best in handling the prediction of children classification, is one of the aims of this study. For that reason, the decision tree (particularly the J48 algorithm), the bayes (the Naïve Bayes Simple algorithm in particular) and neural network classification methods were applied for conducting experiments to build the best model. Summary of experimental result for the three classification algorithms that scores higher accuracy from each is presented in table 4.13.

Classification Model	Overall accuracy (17044 records)	
	Correctly classified	Misclassified
Decision Tree	8422(98.83 %)	100(1.17%)
Naïve Bayes	8379 (98.32%)	143 (1.68%)
Neural Network	8425(98.86%)	97 (1.14 %)

Table 4.14 Accuracy of the J48 decision tree, Naïve Bayes and neural network

The result of the three algorithms are compared each other by their overall classification accuracy (performance). And as can be clearly shown in the table above (table 4.14), the overall performance of the decision tree model was 98.83% with 8422 data sets. However, the classification accuracy of the naïve bayes model with this data size and parameter was 98.32%. In naïve bayes classifier the highest classification accuracy was achieved in the same datasets with decision tree and 50/50% test split option. Furthermore the classification accuracy of the neural network model was 98.86%.

The J48 decision tree has shown second better classification performance after neural network. Hence, it is really reasonable to conclude that the J48 decision tree model is the best classifier model for implementing of children classification applications in the love for children organization. The reason for the J48 decision tree to perform better than neural network is because of the linearity nature of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular children datasets. Regarding the Naïve bayes, scoring a lower accuracy than the J48 decision tree is due to its assumption that each attribute is independent of other attributes. Moreover, in terms of ease and simplicity to the user the J48 decision tree is more self-explanatory. It generates rules that can be presented in simple human language.

Therefore, it is plausible to conclude that the J48 algorithm is more appropriate to this particular case than the Naïve bayes and neural network method. So, the model that is developed with the J48 decision tree classification technique is taken as the final working classification model.

Chapter Five

Conclusion and Recommendation

5.1 Conclusion

Recent advances in communication technologies, on the one hand, and computer hardware and database technologies, on the other, have made it all the more easy for organizations to collect, store and manipulate massive amounts of data. Having concentrated on the accumulation of data, the question is what to do next with this valuable resource? Indeed, the data contains and reflects activities and facts about the organization. The increase in data volume causes great difficulties in extracting useful information and knowledge for decision support. It is to bridge this gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as Data Mining or Knowledge Discovery in Databases (KDD) has emerged in recent years.

The objective of this research undertaking was to explore the possible application of data mining techniques in the orphan and vulnerable children, and particularly to the Love for children NGO organization, by developing a predictive model that could help for the organization internal purpose and for providers to identify children at risk so that they can be treated before the condition becomes something difficult to the children, made use of three predictive modeling techniques, decision tree, neural networks and Bayesian network to address the problem.

As it is often the case, the KDD process was undertaken in phases. The process adopted in this research can be described as constituting five phases: Understanding of the problem domain, understanding of the data, data preprocessing, data mining, and evaluation and interpretation of data mining results.

Understanding of the problem domain: The problem domain was explored to have insight into the area and to be able to define the problem to focus on. During this phase close interaction with the domain experts and review of documents was made to good effect.

Understanding of the data: The discovery task was run on the children related database that consists of 17044 records in more than 20 tables describing a total of 11 attributes.

Data mining: Classification, data mining technique was applied to accomplish the goal of the research. To this effect, the decision tree, Bayesian network and neural network algorithm, which is an implementation of the classification in the Weka 3.7.4 software, was used.

Data preprocessing: Data cleaning and preparation tasks were carried out to handle missing value and noise. Irrelevant, attributes were also excluded in this stage of the research.

Data evaluation and interpretation: In general, encouraging results were obtained by employing Bayesian network, neural networks and decision tree approaches. Although Bayesian network, neural network and decision trees showed comparable accuracy and performance in predicting the children condition, the decision tree approach seems more applicable and appropriate to the problem domain since it provides additional features such as rules can be expressed in human language so that anyone can easily understand how and why a classification of children is made.

5.2. Recommendation

This research work is conducted mainly for academic purpose. However, it is the researcher's belief that the findings of the research will help Governmental and non- governmental organizations to work on the application of data mining techniques to gain competitive advantage in their organization. Moreover, the research work can contribute a lot towards a comprehensive study in this area in the future.

In the course of doing this study and on the basis of the findings of the research work, the researcher has come up with the following recommendations.

- The predictive model, which is developed in this research, generated various patterns. For the company to use it effectively there is a need to design a knowledge base system, which can provide advice for the domain experts.

- The model building process in this investigation was carried out in three algorithms of classification model .that are J48 decision tree, neural network and Bayesian network algorithm. Though, the results were encouraging, further investigation needs to be done using other classification techniques such as Support Vector Machine and other data mining techniques such as clustering and association rule to see if they could be more applicable to the problem domain.

- Although three of the techniques reported promising results and hence could be applied in the area of child classification, decision tree tends to perform better. Hence, it would be more optimal for Love for children organization to employ the model developed with this technique.

References

Adrians P, Zantinge D. (1996). Data mining. Addison-Wesley Longman, England.

Alan R. (1995). Data Mining an Introduction, Version 2.0, Queens University Belfast. http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html

Barbara D, Couto J, Jajodia S. (2001), "a test bed for exploring the use of data mining in Intrusion detection", SIGMOD Record, PP.15-24,

Brachman and Anand,(1996), "The process of Knowledge discovery in database", American Association for Artificial Intelligence Press, PP. 37-57, California.

Denekew, A. (2004) . Application of data mining to support customer relationship management (CRM) at Ethiopian airline. Addis Ababa University, Unpublished master's thesis.

Deshpande, S. and Thakare, V.,(2010),' Data Mining System and Applications: A review, Internation Journal of Distributed and Parallel System (IJDPS), 1(1), 32-44.

Dunham M. Sridhar S. (2006) "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition.

Dunham, M.H. (2003). Data mining introductory and advanced topics: Upper Saddle River, NJ: Pearson Education, Inc.

Elkan, C.1997. Naive Bayesian learning. Harvared University. Viewed 27 February 2011, homepage : http://www.nbl.com.

Fadzilah S. Mansour A.Mining Enrolment Data Using Predictive and Descriptive Approaches, Applied Sciences, College of Arts & Sciences, University Utara Malaysia, Malaysia

Fayyad, U. Piatetsky-Shapiro, G. & Smyth, P. (1996), “From data mining to knowledge Discovery in database”, American Association for Artificial Intelligence Press, cambridge.

Feuerstein M. (1997). Poverty and healthy: Reaping a Richer Harves, Macmillan Education LTD. London.

Global Action for Children (2005), Washington DC : http://www.globalactionforchildren.org/issues/basic_education/] accessed on 1st august, 2009.

Han, J. & Kamber, M. (2001). Data Mining: Concepts and Techniques. San Fransisco; Morgan Kufman Publishers.

Helen T. (2003). Application of data mining technology to identify significant patterns in census or survey data: The case of 2001 child labor Survey In Ethiopia. Addis Ababa University, Unpublished master’s thesis.

Henock, W. (2002). The application of data mining to support customer relationship management at Ethiopian airlines. Addis Ababa University, Unpublished master’s thesis.

Irish Aid, Department of Foreign Affairs (2008) Education Policy and Strategy: Building sustainable Education Systems for Poverty Reduction, Irish Aid, Dublin.

Khalid, S. (2003).Using Decision Trees to predict customer behavior Accessed date: October 12, 2011 http://www.expresscomputeronline.com/20040412/technology01.shtml

Klosgen, W. Zytkow, J (2002) ’’Handbook of Data Mining and Knowledge Discovery’’, Oxford University Press

Kumneger, F. (2006). Application of data mining techniques to support CRM in Ethiopian shipping lines: Addis Ababa University, Unpublished master’s thesis .

LoriBowen, A. (2006). Data Mining for Information Professionals,
http://techessence.info/node/53 .

Leul, W.(2003). The application of data mining in crime prevention: the Case of oromia police Commission. Addis Ababa University, Unpublished master's thesis.

Matheus C. Chan, P. And Piatetsky G.(1993), "Systems for knowledge discovery in databases", IEEE Transactions on Knowledge and Data Engineering 5(6), 903–913.

Meera G., Gandhi and Srivatsa S,(2010) "Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier for an NIDS Environment", Advances in Computational Sciences and Technology, Vol. 3, PP. 291–304.

Mwaanga, S. (2003) the empowerment through participation in women football and EDUSPORT for low SES and HIV/AIDS at risk adolescent girls in Zambia. Unpublished Masters Thesis in Sport and Physical Education, Norwegian University of Sport and P.E, Oslo.

Nanda A.(2010), "Data Mining & Knowledge Discovery in Database: An AI perspective" , Proceedings of national Seminar on Future Trends in Data Mining .

Qi Luo,(2008),"Advanced Knowledge Discovery and Data Mining", IEEE discovery and Data Mining, Workshop on Knowledge Discovery and Data Mining.

Report of the Seminar on Street Children, Organized by the National Children's Commission, June 2, 1988 Addis Ababa .

Roshan R. October 29, 2011, "Difference Between KDD and Data mining", http://www.differencebetween.com/difference-between-kdd-and-vs-data-mining/ ,

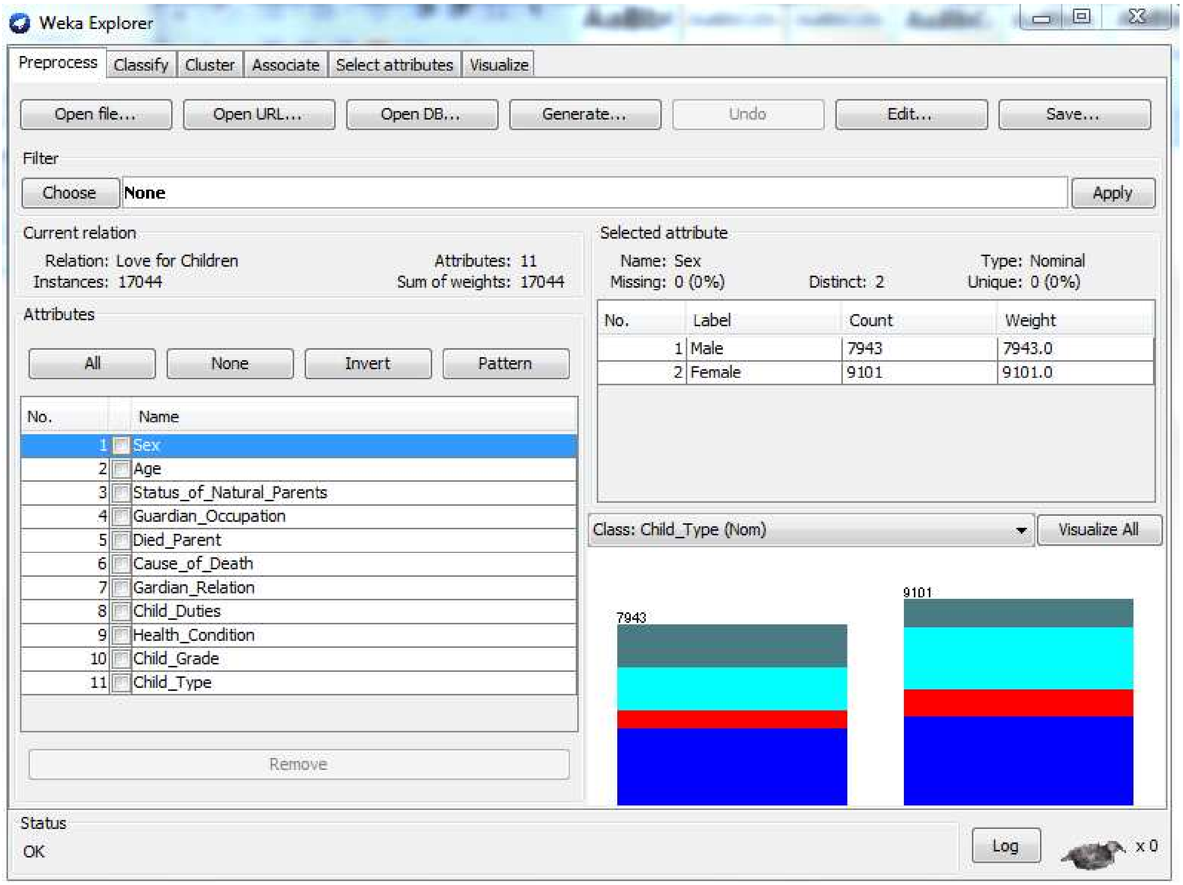
Shegaw, A.(2002). Application of data mining technology to predict child mortality pattern (case of butajira Rural Health Project (BRHP)). Addis Ababa University, Unpublished master's thesis .

Thearling ,K. (2003). An introduction to data mining. Viewed 7 march 2011, Http:// www3.shore.net/~ Kht/text/dmwhite.html UNICEF (2006) the state of the World Children, UNICEF, UNICEF.

Witten, I. H. and E. Frenak.(2005).Data mining : Practical machine learning tools and techniques. Sanfrancisco: Elsevier.

Woldekidan, k.(2003). Application of KDD on crime data to support the Advocacy and awareness raising program of forum on Street children Ethiopia", Department of Information science, M.Sc. Thesis, Addis Ababa University.

Appendix1: The weka 3.7.4 interface with the Dataset Opened to start the first Classification



Filter

None

Current relation

Relation: Love for Children
Instances: 17044
Attributes: 11
Sum of weights: 17044

Selected attribute

Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Male	7943	7943.0
2	Female	9101	9101.0

Attributes

No.		Name
1	□	Sex
2	□	Age
3	□	Status_of_Natural_Parents
4	□	Guardian_Occupation
5	□	Died_Parent
6	□	Cause_of_Death
7	□	Gardian_Relation
8	□	Child_Duties
9	□	Health_Condition
10	□	Child_Grade
11	□	Child_Type

Class: Child_Type (Nom)

Status

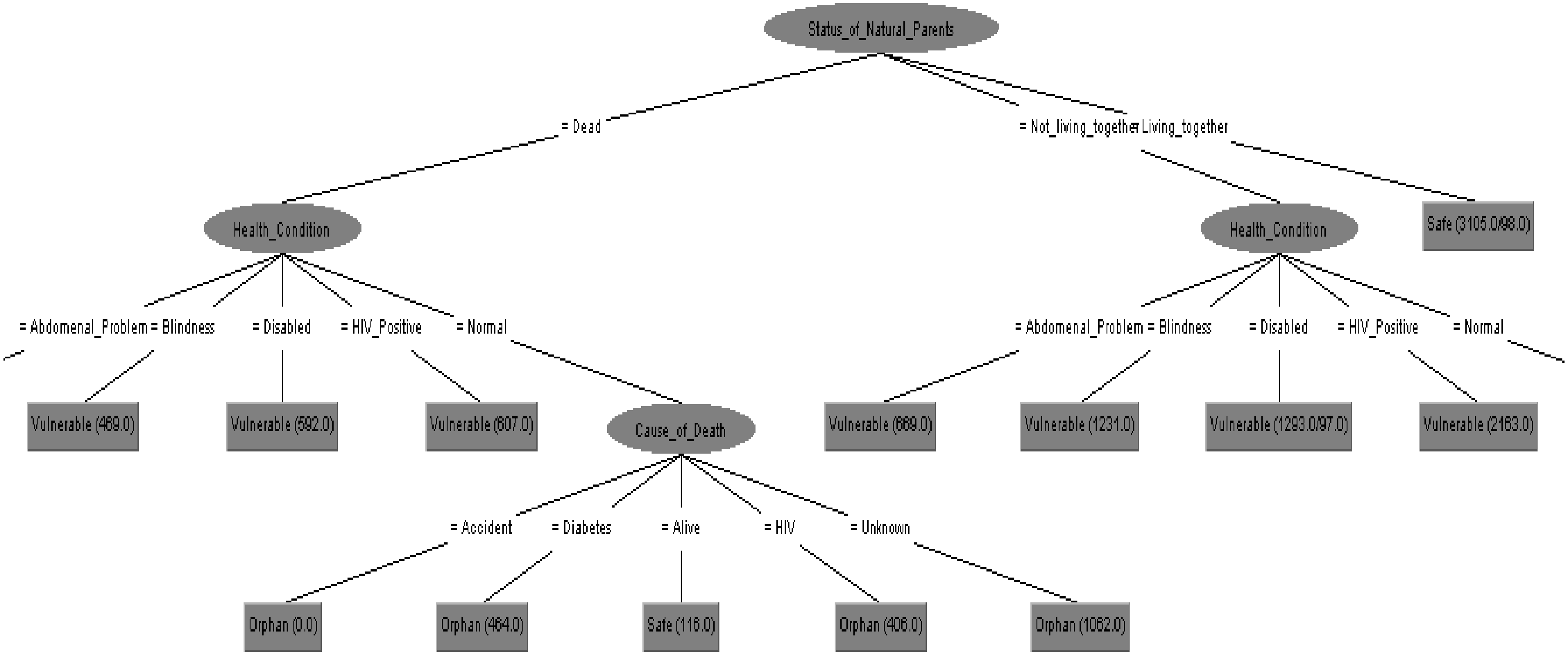
OK

Appendix 2: Sample of the decision tree generated with 50/50 % Validation techniques

```
=== Run information ===
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    Love for Children
Instances:   17044
Attributes:  11
             Sex
             Age
             Status_of_Natural_Parents
             Guardian_Occupation
             Died_Parent
             Cause_of_Death
             Gardian_Relation
             Child_Duties
             Health_Condition
             Child_Grade
             Child_Type
Test mode:   split 50.0% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Status_of_Natural_Parents = Dead
|   Health_Condition = Abdomenal_Problem: Vulnerable (298.0)
|   Health_Condition = Blindness: Vulnerable (469.0)
|   Health_Condition = Disabled: Vulnerable (592.0)
|   Health_Condition = HIV_Positive: Vulnerable (607.0)
|   Health_Condition = Normal
|   |   Gardian_Relation = Father: Orphan (0.0)
|   |   Gardian_Relation = Grand_Mother: Orphan (1403.0)
|   |   Gardian_Relation = Mother: Orphan (0.0)
|   |   Gardian_Relation = Mother_and_Father: Safe (116.0)
|   |   Gardian_Relation = Sister: Orphan (529.0)
```

```
Status_of_Natural_Parents = Not_living_together
|  Health_Condition = Abdomenal_Problem: Vulnerable (669.0)
|  Health_Condition = Blindness: Vulnerable (1231.0)
|  Health_Condition = Disabled
|  |  Child_Grade <= 5: Vulnerable (1066.0)
|  |  Child_Grade > 5
|  |  |  Sex = Male: Vulnerable (36.0)
|  |  |  Sex = Female: Single_Orphan (191.0/94.0)
|  Health_Condition = HIV_Positive: Vulnerable (2163.0)
|  Health_Condition = Normal
|  |  Died_Parent = Both: Orphan (51.0)
|  |  Died_Parent = Mother: Single_Orphan (51.0)
|  |  Died_Parent = Father: Single_Orphan (4467.0)
|  |  Died_Parent = None: Single_Orphan (0.0)
Status_of_Natural_Parents = Living_together: Safe (3105.0/98.0)
Number of Leaves:     20
Size of the tree:     27
Time taken to build model: 0.31 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances        8422             98.8266 %
Incorrectly Classified Instances       100              1.1734 %
Kappa statistic                     0.9832
Mean absolute error                     0.0107
Root mean squared error                   0.0746
Relative absolute error                 3.0728 %
Root relative squared error             17.8874 %
Coverage of cases (0.95 level)           99.3663 %
Mean rel. region size (0.95 level)       26.8863 %
Total      Number      of      Instances                                          8522
```

Appendix 3: the partial overview of decision tree



Appendix 4: Sample rules to predict new instances of records of children's In to their Corresponding class

Rule #1: if Status_of_Natural_Parents = Dead, Health_Condition = Abdomenal_Problem, Blindness, Disabled, and HIV_Positive: Then the child is classifies in **vulnerable**

Rule #2: if Gardian_Relation = Father, Gardian_Relation = Grand_Mother, Gardian_Relation = Mother and Gardian_Relation = Sister: Then the child is classifies in **Orphan**

Rule #3: If Status_of_Natural_Parents = Not_living_together, Health_Condition = Abdomenal_Problem, Health_Condition = Blindness, and if Health_Condition = Disabled and Child_Grade <= 5 and if Child_Grade > 5 and Sex = Male: Then the children is classifies in **Vulnerable**

Rule #4: if Health_Condition = Normal, Died_Parent = Both: Then the child is classifies in **Orphan**

Rule #5: If Died_Parent = Mother, Died_Parent = Father, Died_Parent = None, and if Child_Grade > 5 and Sex = Female: Then the children is classifies in **Single_Orphan.**

Rule#6: If Status_of_Natural_Parents = Living_together: and Gardian_Relation = Mother_and_Father: Then the child is classifies **in Safe**

Declaration

I declared that the thesis is my original work and has not been presented for a decree in any other university and that all sources of material used for the thesis have been duly acknowledged.

Gebremedhin Gebreyohans

June 2012

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

Ato Getachew Jemaneh

June 2012