ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

APPLICATION OF DATA MINING TECHNOLOGY TO PREDICT CHILD MORTALITY PATTERNS: THE CASE OF BUTAJIRA RURAL HEALTH PROJECT (BRHP)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN INFORMATION SCIENCE

BY

SHEGAW ANAGAW

JUNE, 2002

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

APPLICATION OF DATA MINING TECHNOLOGY TO PREDICT MORTALITY PATTERNS: THE CASE OF BUTAJIRA RURAL HEALTH PROJECT(BRHP)

BY

Shegaw Anagaw

Signature of the Board of Examiners for Approval

___________________________	___________________________
___________________________	___________________________
___________________________	___________________________
___________________________	___________________________

DECLARATION

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university

Shegaw Anagaw
June 2002

The thesis has been submitted for examination with our approval as university advisors.

Tesfaye Birru(Ato)

Alemayehu Worku(Dr.)

Dereje Teferri (Ato)

DEDICATION

To my late sister Tiruayehu Anagaw

ACKNOWLEDGEMENTS

My special thanks go to my advisor Ato Tesfaye Biru. Besides bringing the research area to my attention, his direction, keen insight, guidance, and skilful pushes to get me explore had a huge impact both in this research and on my academic development. I also would like to express my heartfelt gratitude to Dr. Alemayehu Worku for his all rounded guidance and support in the course of writing this Thesis. I am also grateful to Ato Dereje Teferi for his comments. I would also like to extend my gratitude to the academic staff members of the community Health Department at Faculty of Medicine, AAU.

My special thanks are also due to my friends Abdulhakim Abubeker, Bahiru Zeinu, Birhanu Kiber, Damtew Miteku, and G/Michael G/Silassie for their all rounded contribution and support in the course of conducting this research work. I would also like to thank my friend in U.S, Ato Gebru Ayalew for his financial and moral support. Thanks is also due to W/ro Ayneabeba Amenu and W/ro Meseret Firew who helped me in typing the manuscript of this thesis. My thanks also go to all the SISA community who have influenced this work in many ways.

Last, but not least, I am grateful to my sweat heart Sofia Luelseghad whose care, tolerance, and attention made the whole process of this research work less difficult than it would be without her.

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AAU: Addis ababa University

BRHP: Butajira Rural Health Project

CART: Classification and regression Trees

CD_ROM: Compact Disk read Only Memory

CHAID: Chi-squared Automatic Interaction Detection

DOS: Disk Operating System

DPT:

EPI6: Epidemiology Version 6

FDRE: Fedral Democratic Republic of Ethioppia

GNP: Gross National Product

ICD-9: International Classification of Diseases, Version 9.

ICU: Intensive Care Unit

KDD: Knowledge Discovery in Databases

OLAP: Online Analytical Processing

PHC: Primary Health Care

RMS: Random Mean Squared Error

SAS

SISA: School of Information Studies for Africa

SPSS: Statistical Packages for Social Sciences

ABSTRACT

Traditionally, very simple statistical techniques are used in the analysis of epidemiological studies. The predominant technique is logistic regression, in which the effects predictors are linear. However, because of their simplicity, it is difficult to use these models to discover unanticipated complex relationships, i.e., non-linearities in the effect of a predictor or interactions between predictors. Specifically, as the volume of data increases, the traditional methods will become inefficient and impractical. This in turn calls the application of new methods and tools that can help to search large quantities of epidemiological data and to discover new patterns and relationships that are hidden in the data. Recently, to address the problem of identifying useful information and knowledge to support primary healthcare prevention and control activities, health care institutions are employing the data mining approach which uses more flexible models, such as, neural networks and decision trees, to discover unanticipated features from large volumes of data stored in epidemiological databases.

Particularly, in the developed world, data mining technology has enabled health care institutions to identify and search previously unknown, actionable information from large health care databases and to apply it to improve the quality and efficiency of primary health care prevention and control activities. However, to the knowledge of the researcher, no health care institution in Ethiopia has used this state of the art technology to support health care decision-making.

Thus, this research work has investigated the potential applicability of data mining technology to predict the risk of child mortality based up on community-based epidemiological datasets gathered by the BRHP epidemiological study.

The methodology used for this research had three basic steps. These were collecting of data, data preparation and model building and testing. The required data was selected and extracted from the ten years surveillance dataset of the BRHP epidemiological study. Then, data preparation tasks (such as data transformation, deriving of new fields, and handling of missing variables) were undertaken. Neural network and decision tree data mining techniques were employed to build and test the models. Models were built and tested by using a sample dataset of 1100 records of both alive and Died children.

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

In this research work, the researcher has proved that an epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities.

CHAPTER ONE

Introduction

1.1 Background

1.1.1 Health Care Practices and The BRHP (Butajira Rural Health Project)

Identifying health-related problems of the community is one of the most important steps in the planning of health care interventions. However, appropriate planning, of health programs in turn depends to a large extent on access to timely and accurate information on demographic characteristics, on the occurrence of major health problems, and on associations with underlying factors.

As Desta (1994) stated community diagnosis requires basic information on demographic characteristics of the population. The availability of valid data on demographic and epidemiological conditions, and on patterns of health care units and their utilization, very much facilitates planning, monitoring, and evaluation of health services. To this end, longitudinal population-based studies are needed to generate sound data on morbidity, mortality, and fertility (Desta, 1994).

Although community-based health information is vital for rational health planning, evaluation, and intervention, one of the fundamental challenges of many developing countries in health care delivery is lack of adequate health information system which collects, compiles, analyses, interprets, and disseminates health related information for planning and decision making. For example, developing countries often lack the most elementary data on mortality where available; they often are not of sufficient quality to provide planners and decision-makers with the required information on levels and trends of mortality.

As Desta(1994) pointed out , in Ethiopia, as many developing countries , current data deficiencies are large and serious in their implications for rational health planning and health research.

Among what should be in place to resolve this challenge is establishing a well-designed epidemiological surveillance system on defined population groups (Desta, 1994). Epidemiological Surveillance represents on-going systems for the collection, analysis and interpretation of health data, which are essential to the planning, implementation and evaluation of health programs. An effective surveillance system also implies timely dissemination of its data to those who need to know. The final link in the surveillance chain is the application of these data to prevention and disease control. Thus, epidemiological surveillance activities are usually included as part of the health information systems of governmental institutions that have responsibility for health care prevention, and disease control (Yemane, et. al., 1999).

The primary purpose of health information is to improve health care delivery. Specifically, at the primary health care level, reliable and timely data about defined population groups can be used as a basis for effective policy formulation and implementation (Desta, 1994). For instance, data collected about child mortality and morbidity can serve to identify major causes and determinants of child mortality and to take preventive actions to reduce the rate of child mortality and to improve child survival. However, in Ethiopia, the capacity to collect, compile, analyze, interpret, and disseminate timely and accurate health information for decision-making is very poor. There is no systematically organized registration of vital events; population based studies are rarely carried out in Ethiopia with the exception of the Butajira Rural Health Project (BRHP) (Desta, 1994).

It was to bridge this gap of access to timely and accurate health related information that the Butajira Rural Health Project was established in 1986 as an epidemiological study (Desta, 1994). The ultimate aim of the project was to provide current epidemiological information system and thereby contribute the improvement of primary health care management and decision-making, especially in the area of primary health care, and particularly at district level. In line with its objectives, the initial tasks for the BRHP during 1986-87 were to perform a census of the population in the selected villages to obtain the baseline population and to establish a system of demographic surveillance with continuous registration of vital and migratory events at a household level. Events registered by the BRHP are birth, death, cause of death, marriage, new household, out-migration, and internal move (Yemane et. al., 1999).

In the BRHP, data are collected monthly by visiting each household. Each household is identified by a unique number within its village, and each individual within their household. The status of each individual within the household is checked during the monthly visit and recorded on a demographic surveillance sheet. Basic demographic, social, and health related data are recorded for each household. Data entry is performed at the Department of Community Health, Faculty of Medicine, AAU in Addis Ababa. Software developed specifically for this purpose using dBase Iv platform, is used to facilitate automatic data checking and entry. The population-based database of the BRHP currently hosts more than 64,000 records about individuals living in the study area. The process of collecting, storing, and analyzing the population-based data gathered by epidemiological studies are facilitated by utilizing modern information and communication technologies.

1.1.2 Data Mining and the Health care

Today, we are witnessing the development of a new chapter in the information revolution caused by the confluence of computers and related technologies. The new technology has drastically changed our society and economy. In information storage and retrieval activities, technology has the potential to realize the ultimate dream of the information retrieval specialist: to make information available to any person, when and where it is required. As Bigus (1996) stated, over the last three decades, the use of computer technology has evolved from piecemeal automation of certain business operations, such as accounting and billing, into today's integrated computing environments, which offer end-to-end automation of all major business processes. Not only the computer technology has changed, but also how that technology is viewed and how it is used in business has changed.

Especially, in the health care sector, technological advancements in the form computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable (Prather et. al., 2002).

Although the capabilities to collect and store data in large computer databases has increased significantly, the relational database technology of today offers little functionality to process and explore data and establish a relationship or pattern among data elements that are hidden or previously unknown (Raghavan, et. al., 1998). Prather et. al. (2002) on his part wrote that , although health care databases have accumulated large quantities of information about patients and their medical conditions, there are only few tools to evaluate and analyze this clinical data after it has been captured and stored. The authors further stated that evaluation of stored clinical data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management.

As Fayyad, Paitesky-Shapiro, and Smyth (1996) wrote, the traditional method of turning data into knowledge relies on manual analysis and interpretation. The writers further stated that, for example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health-carte organization; this report will be used for future decision making and planning for health-care management.

Specifically, as data volumes grow dramatically, data analysis based on manual methods is becoming completely impractical in many domains. Prather et. al. (2002) argued that to evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data and to discover new patterns and relationships hidden in the data. It is due to these challenges of searching for knowledge in relational databases and our inability to interpret and digest these data as readily as they are accumulated, which has created a need for a new generation of tools and techniques for automated and intelligent database analysis. Consequently, the discipline of knowledge discovery in databases (KDD), which deals with the study of such tools and techniques, has evolved into an important active area of research (Raghavan, et. al., 1998).

According to Larvac(1998), in the health care sector, the widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer assisted analysis. Such an extensive amount of data gathered in medical databases require specialized tools and methods that can be used to discover new information and knowledge which is useful in decision making and problem solving.

Faced with the tremendous economic and competitive pressures, the health-care industry has started to mine its data to cut costs, improve quality and save lives. In support of this notion, Bresnahan (1997) argued that one way in which data mining is helping health-care providers cut costs and improve care is by showing which treatments statistically have been most effective. For example, once hospital administrators recognize that stroke patients are less likely to develop respiratory infections if they can swallow properly, they can educate their physicians and institute a standard policy to identify and provide therapy to those who have difficulty of swallowing.

The data mining process which serves as a means of searching previously unknown, actionable information from large databases can also be used to improve the quality and efficiency of care of patients which is known in the health care industry as “outcomes measurements.” Outcomes measurement involves examining clinical encounter information, insurance claims and billing data to measure the results of past treatment and process (Bresnahan, 1997).

Bresnahan(1997) further stated that since many of the issues and problems associated with outcomes measurement apply to data mining, health care providers can identify areas of improvement or capitalize on successful methods. Outcomes measurement using data mining also helps health-care institutions to evaluate their doctors and facilities. Physicians and hospitals benefit from knowing how they compare with their peers, and the parent company saves money by getting all of its employees up to par. Along the same lines, outcomes measurement by using data mining technology also lets caregivers identify people statistically at risk for certain ailments so that they can be treated before the condition escalates into something expensive and potentially fatal.

Data mining and knowledge discovery can also “close the loop” between clinical data mining capture and evidence-based decision support by facilitating the conversion of clinical data into evidence for future decision (Downs and Wallace, 2001).

Recently, data mining techniques have been also applied for survival analysis and the prediction of prognosis in cancer treatment (Theeuwen, kappen, and Neijt, 2001). Plate et. al. (1997) also described that although epidemiological data is traditionally analyzed with very simple statistical techniques, today, flexible models, such as neural networks, have the potential to discover unanticipated features in epidemiological data.

Prather et. al. (2002) have also used the techniques of data mining to search for relationships in a large clinical databases. Specifically, the authors used a dataset accumulated on obstetrical patients and evaluated for factors potentially contributing to preterm birth using exploratory factor analysis.

While the application and utilization of data mining technology in the health care sector is steadily growing fast in the developed world, its applicability remains to be unknown in the Ethiopian health care sector. Given experiences elsewhere in terms of the benefits acquired in applying data mining technology in the health care sector, it is only proper to explore the relevance and potential advantage of such a state of the art technology in the Ethiopian health-care context. Thus, in this research project, since there is high child mortality rate at a national level in general and at the BRHP study area in particular, the researcher is motivated to assess the potential applicability of data mining technology to predict child mortality patterns in rural Ethiopia by using an epidemiological database. For reasons of familiarity and availability of electronic data, the researcher chose the Butajira Rural Heath Project (BRHP) to conduct the study.

1.2 Statement of the Problem and its Importance

The underlying research problem that necessitated this research is the existence of high infant and child mortality at a national level in general and at the Brtajira project area in particular. Specifically at the district of Butajira, although data on child mortality and morbidity is periodically gathered by the BRHP epidemiological study, due to lack of appropriate data analysis tools this data is not practically used to alleviate the problems faced by health-care professionals, planners and policy makers to identify major determinants of infant and child mortality in order to plan and implement age-specific intervention strategies to reduce infant and child mortality in rural Ethiopia.

In Ethiopia, mortality figures indicate very high infant and child mortality rates in rural areas (Desta, 1994). Although the extension of health services to the underserved rural populations has been an explicit goal in national development plans in Ethiopia over the last two decades, the rural community is still disfavored in terms access to health services. In the Butajira district, residents still suffer from preventive diseases. As of a research report made by Desta (1994), the leading causes of child mortality in the BRHP study area are ARI (Acute Respiratory Infections) and diarrhea.

According to a recent report made by the FDRE Central Statistical Authority, immunization coverage among children under 5 years of age is less than 50 percent against Measles, and DPT in rural areas while the coverage is 84 percent in urban areas.

On the other hand, prevalence of diarrhea/fever among the under five years children stands at about 27 percent with rates being slightly higher among rural children than urban irrespective of the level of expenditure group of households. Assessment of accessibility of health services in terms of distance to the nearest health center indicates that rural households are disfavored. That is, about 94% of urban households can get health facility with a distance of 5 kilometers compared to one-third of rural households that have access to health services within that distance. The report also revealed that prevalence of illness is higher among rural population, young children and aged individuals (FDRE Central statistical Authority on the Year 2000 Welfare Monitoring Survey, 2001).

Thus, to alleviate the current problem of high infant and child mortality at a national level in general, and among rural communities in particular, developing a capacity to design and implement an effective health information system that can provide timely and accurate health information on defined population groups is very crucial. Specifically, timely and reliable data collected about child mortality patterns can be used to identify major determinants and risk factors for child mortality and to take preventive actions so that to improve child survival in the region.

Information about the major causes of infant and child death is definitely of greater importance for age-specific health care intervention, planning and for the improvement of child survival. As Desta (1994) pointed out, the starting point for setting priorities for child survival is to assess the occurrence of easily identifiable causes of death in childhood. Identifying age-specific mortality patterns for common causes of death could serve as a basis for selective intervention strategies. Last and kandel (2002) also wrote that causes of death can have a significant impact on the budget priorities of the Health Ministry. Higher frequency of certain cause in a defined population group (defined by gender, age, origin etc.) can increase the budget, aimed at the death rate of that group. The portion of a cause in the entire population may be used as another important criterion: preventing the leading causes of death should obtain a higher priority.

However, as Desta (1994) stated developing countries lack the most elementary data on mortality patterns. Those available are often not sufficient quality to provide health care planners with required information on levels and trends of child mortality. In Ethiopia, the practical challenge for health care providers, planners, and policy makers working in primary health care prevention and control activities is lack of timely and reliable health information on the health states of defined population groups.

As it is described in the background section of this chapter, the BRHP was established to provide an epidemiological information system and thereby contribute to improved management and decision-making, especially in the area of primary health care, in particular at district level. According to Yemane et. al. (1999), the population based data that has been collected for the last 15 years by the BRHP would serve as a source of data to assess fertility and mortality trends, to analyze public health and behavioral determinants of mortality and morbidity as well as to assess coverage and utilization of health services in the area in relation to health needs of the people.

By using the information gathered by the BRHP epidemiological surveillance system, several studies were conducted (Yemane, et. al., 1999). Specifically, previous studies conducted on infant and child mortality indicate the existence of very high rate of infant and child mortality in the BRHP study area. Those studies have also reported that, Diarrhea and Acute Respiratory infection (ARI) are the leading causes of infant and child mortality and morbidity in the study area. The factors identified in those studies as determinants of these disease entities are maternal education and occupation, duration of breast-feeding, place of residence, household income, infant's birth weight, and the birth order interval (Yemane et. al., 1999).

However, the problem is that all those previous studies were conducted by using a very small proportion of the database. Besides, in those studies, data analysis was conducted by using simple statistical techniques (such as regression and verification techniques). Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world databases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases (Plate et. al., 1997).

Particularly, in order to identify patterns in infant and child mortality and the associated risk factors, it is difficult to conclusively attribute the problem to a certain set of factors since the parameters (attributes or features) involved are too many. In relation to this, Last and Kandel (2002) described that the tools used in the research on death causality have been so far limited to the statistical techniques, like summarization., regression, analysis of variance, etc. However, such methods of data analysis have concentrated on identifying the leading causes of death and testing the association of single factors with certain diseases. No complex models (involving more than one factor) have been built.

In addition, although the BRHP epidemiological study has collected and stored large volume of population-based data for the last 15 years, the accumulated data is not fully utilized to support primary health care prevention and control activities in the region. In such situations, new information technologies like data mining may be helpful in discovering hidden data patterns. They may be used to build models that can support child health care prevention and control activities in the region.

Thus, in this study, the researcher investigated how social, economical, behavioral, environmental, and health related factors affect child mortality at the district of Butajira by applying the new computerized methods of data mining technology.

The primary goal of this research work was thus, to assess how well a predictive model built by using data mining techniques would perform in predicting the risk of mortality among children based solely upon demographic, parental, environmental, and epidemiological history factors, without using diagnostic tests or physical exam findings. This type of prediction model might have an application outside of clinical settings to support health care workers in taking preventive actions to reduce child mortality in the region, as well as to assist health care planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at improving child survival in the region. Thus, the results of this data mining process can be used to distinguish children at high risk of mortality from those with low risk of mortality.

The hypothesis in which this research is based on is that if the existing data in the BRHP database is analyzed by using data mining techniques and tools, it can provide knowledge that can facilitate the improvement of child survival by identifying significant and meaningful patterns among different data elements which were hidden or previously unknown. Specifically, in relation to the problem domain of this research project, if the right data (i.e., related to children) included in the surveillance system is selected, organized, cleaned, and analyzed using data mining tools and techniques it might show a result that could help to develop a model to predict the risk of mortality among children and to identify major determinants of child mortality in rural Ethiopia.

1.3 Objectives of the Research Undertaking

1.3.1 General Objective

The general objective of this research is to investigate the potential applicability of data mining technology in developing a model that can support primary health care providers, policy makers, planners etc. to identify the major determinants of child mortality and to prevent and control child mortality at the district of Butajira.

1.3.2 Specific objectives

In order to achieve the specified general objective, the research work has undertaken the following specific objectives:

- Conduct a through review of literature on the existing data mining techniques and methods in general, and their application in the health-care sector in particular.
- Assess different neural network and decision tree data mining application software that are more appropriate to the problem domain, and select the best software.
- Select and extract the data set required for analysis from the database of Butajira Rural Health Project;
- Data making or preparing the data for analysis which includes adjusting inconsistent data encoding, accounting for missing values, and deriving other fields from existing ones;
- Build and test models using BrainMaker and See5 software;
- Report results and make recommendations.

1.4 Research Methodology

In this study, in order to assess how well a predictive model built by using data mining techniques would perform in predicting the risk of mortality among children based solely upon demographic, parental, environmental, and epidemiological history factors, the researcher has used the methodology suggested by Berry and Linoff (1997). This methodology assumes that the business problem has already been identified and hence directly proceeds to the different data mining steps that need to be carried out in order to develop a model for the data-mining project.

So, based on the suggested methodology, for this specific data-mining project, the following steps were undertaken:

A. Identifying Sources of Pre-classified Data

In order to identify sources of pre-classified historical data collected about children by the BRHP, the researcher had conducted subsequent discussions with the system administrator of the BRHP and staff members of the Community Health Department of AAU. As a result, the researcher identified two main sources of data stored in electronic format. The first source of data identified was the main database of the project. This is a database, which is updated every month and incorporates every new information gathered about the study population through census and the surveillance system. This database stores large amount of community based information gathered for the last 15 years about the demographic, socio-cultural, and health states of all individuals living in all the ten selected villages of the BRHP study area.

The second source of data identified was the separate ten years' (1987-1996) surveillance data. This is a definitive version of the main database, but it was purposefully cleaned and prepared to provide a separate dataset for those who

are interested for the analysis of the population-based data gathered by the BRHP. This separate ten years' data set contains a total of 64,077 records of individuals registered in all the ten villages of the BRHP study area.
Since this ten years' data set is a dataset created to provide a separate data set for those who wants to work with, it was this dataset that was used as a source of data required for this data mining research project.

Thus, for this research project, the target dataset required to build a model that predicts the risk of child mortality was selected and extracted from this ten years' surveillance dataset. The data set selected for this research project consists of all children who were born between January 1, 1987 and December 30, 1996 in all the ten villages of the Butajira study area. Throughout this period a total of 15,667 live births were registered. Of these records of live births, a total of 2,330 records were about children who were born live, but who died after some time. So, to build predictive models using neural network and decision tree techniques, a sample dataset consisting of 1,100 records from the two classes of children (i.e. alive and died) was selected randomly from this target dataset.

B. Preparing Data for Analysis

Data to be mined will be collected in a new database. This will help to apply data mining tools and algorithms on the data. As such, the relatively clean data that resides in the database will be refined and processed before it undergoes the data mining process. However, this process of data cleaning and pre-processing is highly dependent on the data mining technique to be used and the particular data mining software to be employed. The data mining techniques used for this research project were neural networks and decision tree techniques. Therefore, different neural network and decision tree software were examined by taking into consideration their application to the problem domain.

Thus, in this research project, data preparation was conducted by considering the requirements of *BrainMaker* (neural network) and *See5* (decision tree software. At this stage, data cleaning and pre-processing tasks like decoding of inconsistent data encoding, handling missing values, and deriving new fields from the already existing ones were performed by taking in to consideration the requirements of the chosen software.

C. Build and Train the Model

Although the choice of data mining techniques for classification tasks seems to be strongly dependent on the application, the data mining techniques that are frequently employed for classification tasks are neural networks and decision trees. As it is indicated previously, for the purpose of this research work the researcher experimented the potential applicability of data mining technology in developing a model that predicts the risk of child mortality at the district of Butajira. To this end the researcher has employed and tested the applicability of neural networks and decision tree techniques to the problem domain.

Brainmaker software was employed to build neural network models. This software partitions the data set prepared for analysis into training and test facts where training facts are used to train and build the models and test facts are used to test the performance of the model. By default the software automatically sets aside 10% of the prepared dataset for testing purposes. In this research, numerous networks (models) were built by using this software, and the performance of those models was tested by using test cases set aside by the software.

Using the data set used to build BrainMaker models, performance of See5 decision tree software was also tested. However, before the dataset was used by See5, it was prepared into a form that is suitable for this specific software. Several models were built by varying the options provided by the software.

1.5 Scope and Limitation of the study

The scope of this research is to appraise the potential applicability of data mining technology in supporting child health prevention and control activities at the district of Butajira. While findings of this research work can fairly be considered as relevant in appraising the potential applicability of data mining technology in the Ethiopian Health Care sector at large, the scope of the current experimental research undertaking is strictly limited to appraising the potential applicability of data mining technology to support primary health care activities at the BRHP study area.

The time that was given to undertake this research work was a serious limitation in developing an application model based on the entire records of children identified in the BRHP main database. Besides, obtaining the data mining software needed for this research work was a challenging task. Particularly, the efforts made to search and select more appropriate and affordable data mining software that can be used to build and test models so that to asses the applicability of data mining technology on epidemiological data sets was rather time consuming.

Another limitation of this study was lack of related literature particularly in relation to the application of data mining technology on epidemiological datasets. Lack of prior experience to data mining technology as well as to the problem domain were also the limitations faced by the researcher.

1.6 Organization of the Thesis

This thesis is divided into five chapters. The first chapter is an introduction part, which contains background to the research work, statement of the problem addressed, objective of the research, and methodologies adopted for the study.

The second chapter deals about data mining technology, methods/techniques used, and its application in the health care sector.

The third chapter is devoted to give further understanding about epidemiological studies in general, and the existing epidemiological surveillance activities of the BRHP in particular. In this chapter, issues related to data collection methods, data quality assurance techniques, data entry and database structure of the BRHP are addressed.

The fourth chapter provides discussions about the different data mining steps that were undertaken in this research work. This includes data collection, data preparation, model building and testing results obtained by using BrainMaker and See5 software.

The last chapter is devoted for the final concluding remarks and recommendations forwarded based on the research findings.

CHAPTER TWO

DATA MINING TECHNOLOGY

In this chapter, an attempt has been made to review the literature on the concepts and techniques of data mining in general and its application in the health care sector in particular which is aimed to provide background about the models to be built.

2.1 what is Data Mining?

It is estimated that the amount of information in the world doubles every 20 months; that is, many scientific, government and corporate information systems are being overwhelmed by a flood of data that are generated and stored routinely, which grow into large databases amounting to giga (and even tera) bytes of data (Deogun, et. al., 2001). The authors further argued that given certain data analysis goal, it has been a common practice to either design a database application on on-line data or use a statistical (or analytical) package on off-line data along with a domain expert to interpret the results. Even if one does not count the problems related with the use of standard statistical packages (such as its limited power for knowledge discovery, the need for trained statisticians and domain experts to apply statistical methods and to refine/interpret results, etc.), one is required to state the goal and gather relevant data to arrive at that goal. Consequently, there is still strong possibility that some significant and meaningful patterns in the database, waiting to be discovered, are missed (Deogun, et. al., 2001).

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. Rea (2001) wrote that the past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. 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. But the data's hidden value, the potential to predict business trends and customer behavior, has largely gone untapped (Levin and Zahavi, 1999, Rogers, 2001, Larvac, 1998).

Thus, to provide useful information and knowledge about a business by going beyond the data explicitly stored, the data stored in databases or paper files should be analyzed and interpreted into knowledge. However, statistical theory and practice, which for many years has been the traditional method to study and analyze data, fail when it comes analyzing large amounts of data (Levin and Zahavi, 1999, Larvac, 2001).

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.

According to Han and Kamber (2001) the major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research (Han and Kamber, 2001).

Data mining is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, heuristics, data acquisition, data visualization, optimization, information retrieval, high end computing, and others (Levin and Zahavi, 1999, Han and Kamber, 2001).

Simply stated data mining refers to extracting or “mining” knowledge from large amounts of data (Han and Kamber, 2001). Cabena *et. al*. (1998) defined data mining as “a new generation of computerized methods for extracting previously unknown, valid, and actionable information from large databases and then using this information to make critical business decisions”.

Rea (2001) wrote, “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”.

Indeed, data mining technology has become a new paradigm for decision making, with applications ranging from database marketing and electronic commerce to fraud detection, credit scoring, warranty management, even auditing data before storing it in a database (Levin and Zahavi, 1999).

2.1.1 Data Mining and Other Statistical Tools

Analyzing data being stored either in electronic or paper form is not a new task. Statisticians have been collecting and analyzing data for ages (Last and Kandel, 2002). The computer itself has been used as a tool for data analysis at least twenty years ago. However, the information age has brought some substantial changes to the area of data analysis. As Last and Kandel (2002) puts it many business processes have been computerized which resulted to a significant increase in the amount of available data.

As Levin and Zahavi (1999) described, when a small data set is involved with only several predictors, one can manipulate the data set manually using statistical methods to search for the combination of predictors and their transformations that best fit the data, according to some statistical criteria. Rea (2001) also argued that to detect unusual patterns and explain patterns using statistical models such as linear models, analysts have used statistical analysis systems such as SAS and SPSS.

According to Last and Kandel (2002) most methods of the classical statistics are verification oriented which are based on the assumption that the data analyst knows a single hypothesis (usually called the null hypothesis) about the underlying phenomenon (Last and Kandel, 2002). In such statistical methods the objective of a statistical test is to verify the null hypothesis. More sophisticated uses of hypothesis testing include one-way and two-way Analysis of Variance (ANOVA), where one or two independent factors are tested for affecting another variable, called the "response". As Last and Knadel (2002) puts it "the hypothesis testing can be a practical tool for supporting a decision-making process, but not for improving our knowledge about the world".

The regression methods (simple linear, multiple linear and non-linear models) represent the more discovery-oriented approach of the classical statistics, because they enable to find the unknown coefficients of mathematical equations relating a dependent variable to its predictors. Regression methods are very efficient in computation but they are limited to use with continuous (numeric) attributes only (Last and Kandel, 2002). Moreover, regression methods assume a pre-determined form of functional dependency (e.g. linear) and provide no indication on existence of their functional dependencies in data.

Unfortunately, statistical theory and practice, which for many years has been the traditional method to study and analyze data, fail when it comes to analyzing large amounts of data (Levin and Zahavi, 1999). A marginally better situation is encountered with the OLAP (Online Analytical Processing) tools, which can be termed visualization driven since they assist the users in the process of pattern discovery by displaying multi-dimensional data graphically. Online Analytical

Processing (OLAP) is the application of traditional query-and-reporting programs to describe and extract what is in the database. The user forms hypothesis about the data relationships and employees OLAP to verify the hypotheses with queries of the database (Trybula, 1997).

As a solution to the limitations observed with traditional statistical methods, the machine learning methods (originally developed to deal, mainly, with the problems of pattern recognition) have been introduced into the data-mining field (Last and Kandel, 2002). Levin and Zahavi (1999) have also stated that up until recently, the ability to analyze and understand volume of data lagged for behind the capability, to gather, store and manipulate the data. But not any more after the advent of data mining technology.

However, it does not mean that data mining has replaced other statistical methods such as OLAP, Regression, etc. Rea (2001) wrote that "statistics have a role to play and data mining will not replace such analysis but they can act upon more directed analysis based on the results of data mining". Graettinger (1999) also stated that "data mining does not replace but rather complements and interlocks with other decision support system capabilities such as query and reporting, on-line analytical processing (OLAP), data visualization and traditional statistical analysis".

2.2 The Data Mining (knowledge Discovery) Process

Basically data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. Describing some of the major stages/steps involved in data mining, Levin and Zahavi (1999) stated that data mining is more than just applying software, it is a process that involves a series of steps to preprocess the data prior to mining and post processing steps to evaluate and interpret the modeling results. According to these authors the process of building and implementing a data mining solution is known as Knowledge Discovery in Databases (KDD). Mannila (2002) argued that discovering knowledge from data should therefore be seen as a process containing several steps like understanding the domain, preparing the data set, discovering patterns (data mining), post processing of discovered patterns, and putting the results into use. The following figure shows a summary of the various processes involved in the process of knowledge discovery in databases.

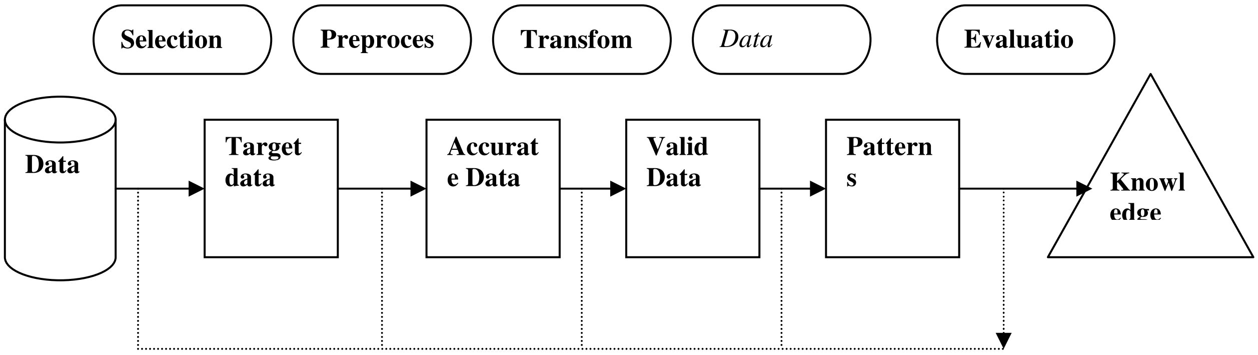


Figure1: Steps in Knowledge discovery In Databases(KDD) process (Based on Fayyad et. al., 1996).

The first step of any data mining process i.e. a clear definition of the business problem involved and the objective function, is an important step to discover useful knowledge for decision-making. Two crows corporation (1999) wrote that to make the best use of data mining you must make a clear statement of your objectives. Without clear understanding of the problem to be solved, it is difficult to identify, select, and prepare the data for mining, or to correctly interpret the results.

The next step involves selecting the target data set for analysis according to some criteria. Extracting the target data to analyze in a way that is consistent with the business problem involved and the objective of the project is the main ingredient of data mining activity since the success or failure of the data mining process highly depends on the data. Regarding the method used to extract the target data, Levin and Zahavi (1999) stated, “often, one can use subjective judgment to extract the relevant target set.” In many cases, one may have to use segmentation analysis, which may require the use of data mining model, such as clustering, to extract the target data set to participate in the data mining process.(Levin and Zahavi, 1999).

The data-cleaning phase (sometimes called scrubbing data) refers to the process of reviewing the data to find incorrect characters or mistransmitted information (Trybula, 1997). Data preparation and preprocessing is often the most time consuming task of the KDD process, especially if data is drawn directly from the company’s operational databases rather than from a data warehouse (Levin and Zahavi, 1999). Extending his description, Rea (2001) stated that, at this stage the data is reconfigured to ensure a consistent format, as there is a possibility of inconsistent formats.

At this stage, you are trying to ensure not only the correctness and consistency of values but also that all the data you have is measuring the same thing in the same way (Two crows corporation, 1998). As Levin and Zahavi (1999) puts it “for data mining purposes, one needs to understand the data, identify key predictors, trace non-linear relationships between data elements, point out important interactions, etc. This requires slicing and dicing the data in different directions, and running a series of frequency and cross tab tables to interrogate the data”. Generally speaking, preparing the data is a step where much time is devoted. Saarenvitra (2001) estimates that this step can take up to 80% of the total project effort.

The other important step in data mining process is transformation of the data. Levin and Zahavi (1999) described “often, the predictive power of data resides in transformation of the data, rather than in the raw data itself.” During transformation the data is made useable and navigable. The data are transformed or consolidated into forms appropriate for mining.

Data transformations are designed to account for non-linear relationships between the dependent variable and one or more independent variables (assuming all the others are constant), identifying pair-wise interaction, perhaps even higher-order interactions, between independent variables, tracking seasonal and time related effects, even transforming data to make them compatible with the theoretical assumptions underlying the model involved (Levin and Zahavi, 1999).

The data mining (pattern discovery) phase, which comes after data transformation step, is the step that invokes data mining models and tools to interrogate the data and convert it into a knowledge for decision making. As Levin and Zahavi (1999) put it those models can be selected from a wide range of models to suit the business issue concerned.

According to Fayyad et.al. (1996), the data mining stage of KDD process consists of two main types of data mining methods: verification oriented (the system verifies user's hypothesis) and discovery oriented (the system finds new rules and patterns autonomously). The discovery-oriented methods can be further partitioned into descriptive (e.g. Visualization) and predictive (like regression and classification).

Interpretation of results is another important stage of the knowledge discovery process. According to Maimon et.al. (2002), at this stage of the knowledge discovery process, the discovered patterns should be presented to a user in an understandable manner. Similarly Levin and Zahavi (1999) argued that evaluation and interpretation of knowledge discovered by the modeling engine is essential so that to make sure the resulting model is any good, and to convert the model results into useful knowledge for decision making. The authors added that the task of knowledge evaluation is often conducted by means of statistical measures and tools such as test of hypothesis, correlation analysis, likelihood function, R^2 measures, misclassification rates, and the like.

In interpreting the model results one should beware of two pitfalls over fitting and under fitting. Over fitting pertains to the phenomenon, that often plagues large scale predictive models, where one gets a very good fit on the data which is used to build the model, but poor fit when the model is applied on a new set of observations (Levin and Zahavi, 1999). To test over fitting, it is necessary to validate the model using a different set of observations than those used to build the model. Under fitting is related to a wrong model that is not fulfilling its mission (Levin and Zahavi, 1999). Under fitting of models could happen due to various reasons like: wrong models, weak data, wrong transformations, missing out the influential predictors in the feature selection process, biased samples, and others. Levin and Zahavi (1999) wrote that, there is no clear prescription to resolve the under fitting issue. Rather the process may require quite some creativity and ingenuity.

Finally, the model that is developed using a data mining technique would be deployed i.e. the mathematical models would be implemented into operational systems (Saarenvitra, 2001). The last step of data mining (knowledge discovery) process is then to use the deployed model to achieve improved results to the problem domain identified at the initial stage of the process.

However, since data mining operation is an iterative process, the above steps are not always to be followed strictly in that order. As Skalak (2001) puts it, each step can conceivably return to any previous one.

2.3 Data Mining Techniques

Data mining can be performed using either a top-down or bottom-up approach. Bottom-up data mining analyses raw data in an attempt to discover hidden trends and groups, whereas the aim of top-down data mining is to test a specific hypothesis (Han and Kamber (2001), 1995). Data mining may be performed using a variety of techniques, including classification, estimation, clustering, etc.

Since the potential number of patterns presented in a real-world database (containing, at least, tens of attributes and thousands of records) may be nearly infinite, most data mining techniques are concealed with a computationally - efficient enumeration of patterns to discover the most useful patterns for the user or task. An important aspect of the mining task lies in the need to extend known techniques and tools in a way that they are robust enough to handle the characteristics of real-world databases (Raghavan; Deogun, and Sever, 2002). The quality of the rules and hence the knowledge discovered is heavily dependent on the algorithms used to analyze the data. Thus, central to the problem of knowledge extraction are the techniques/methods used to generate such rules.

The core of an algorithm constitutes the model upon which the algorithm is built (Raghavan, Deogun, and Sever, 2002). As such choosing the appropriate model, realizing the assumptions inherent in the model and using a proper representational form are some of the factors that influence a successful knowledge discovery. Thus, evaluating and selecting appropriate model for a given knowledge discovery task is essential.

Data mining methods can be classified by the function they perform, (i.e the kind of patterns to be found in data mining tasks) or according to the class of application they can be used in. According to Han and Kamber (2001), data mining tasks can be classified into two categories: predictive and descriptive.

In predictive modeling tasks, one identifies patterns found in the data to predict future values (Levin and Zahavi, 1999). Predictive modeling consists of several types of models such as classification, regression and AI-based models. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. Where as descriptive techniques are sometimes referred to as unsupervised learning since there is no already-known result to guide the algorithms (Two crows corporations, 1999).

On the other hand, descriptive models belong to the realm of unsupervised learning. Such models interrogate the database to identify patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, among others, belong to this family of descriptive models (Levin and Zahavi, 1999; Han and Kamber, 2001).

2.3.1 Classification Methods

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. However, the quality of the discovered knowledge is heavily dependent on the algorithms used to analyze the data (Deogun, et. al., 2001).

Classification methods create classes by examining already classified data (cases) and inductively finding the pattern (or rule) typical to each class (i.e., learning is supervised) (Levin and Zahavi, 1999). These rules support specific tasks and are generated by the repeated application of a certain technique, or more generally an algorithm, on the data (Deogun et. al., 2001). Thus, central to the problem of knowledge extraction are the techniques /methods used to generate those rules/patterns. The core of an algorithm constitutes the model upon which the algorithm is built (Deogun, et. al., 2001). Han and Kamber (2001) describing the classification task stated that data classification is a two-step process. In the first step, a model is constructed by analyzing database tuples described by the attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. In the second step the model is used for classification. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. Then the accuracy of a model on a given test set is evaluated. As Rea (2001) puts it once classes are defined the system should inter rules that govern the classification therefore the system should be able to find the description of each class. The writer further explains that the descriptions should only refer to the predicting attributes of the training set so that the positive examples should satisfy the description and none of the negative. A rule is said to be correct if its description covers all the positive examples and none of the negative examples of a class. Basically, in classification tasks, the system, given a case or tuple with certain known attribute values should be able to predict to which class this case belongs to.

For this research project, a classification task is to be carried out since a model is to be built by using the pre-classified data of past records of children that are included in the study base of the BRHP.

Although the choice of techniques suitable for classification tasks seems to be strongly dependent on the application, the data mining techniques that are frequently employed for classification tasks are neural networks and decision trees (Plate, et. al., 1997). So, since this research project is a classification task conducted by using these techniques, they are discussed below under subsections 2.3.1.1 and 2.3.1.2.

2.3.1.1 Neural Networks

A Neural Network is an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning (Rea, 2001).
Neural networks developed due to difficulty in creating the basic intelligence in computers using the conventional algorithmic approach (Pudi, 2001). Conventional computers are good at following explicit instructions over and over again (Berry and Lionff, 1977). Stergious and Siganos also put the conventional computers take an algorithmic approach where the computer has to follow steps of instructions in solving a problem. Using algorithmic problem solving approach scientists have been able to create machines that can solve complicated logical and mathematical problems. However, with the algorithmic approach, it proved difficult to create a machine that had general human intelligence. By general intelligence we mean every day tasks such as recognizing a face, recognizing a speech, making a cap of coffee etc. (Grove, 1996).

The challenge for scientists to create intelligent machines encouraged artificial intelligence workers to consider the structure of the brain. 'To understand human intelligence and make programs that perform in an intelligent way we must copy the structure of the brain. This is the basic idea behind rural networks' (Grove, 1996).

Neural network, just like the brain, 'is composed to large number of highly interconnected processing elements working in parallel to solve a specific problem' (Siganos, 1996). These networks have the capacity to learn, memorized and create relationships amongst data.

Neural networks, unlike the conventional algorithmic computers, cannot be programmed to perform specific task. Neural networks learn from examples, rather than being told rules or mathematical formulas (Lawrence, 1994). For instance, to generate a model that performs sales forecast neural network only needs to be fed raw data related to the problem. The raw data could include past sales, prices, competitor's prices, and other economic variables. The neural networks then learn from these facts and products a model that can be used to provide prediction of future sales when provided with the independent variables (Z Solutions, 1999).

However, 'Neural networks and conventional algorithmic computers are not in competition but complement each other. There are tasks more suited to an algorithmic approach like arithmetic operations and tasks that are more suited to neural networks' (Stergiou and Siganos, 1996). The writers also add that there are tasks that are best handled by combining the two approaches.

2.3.1.1.1. Brief History of Neural Networks

The original work on neural networks started even before the emergence of digital computers i.e. in the 1940s (Berry and Linoff, 1997). In 1943 McCulloch and Pitts made the first model of the biological neuron. The model described neuron as 'linear threshold computing unit with multiple inputs and a single output of either 0, if the nerve cell remains inactive, or 1, if the cell fires' (Fraser, 2000). The further discusses that the neuron fires if the sum of the inputs exceeds a specified threshold.

In the 1950s, when digital computers became available, scientists implemented models called perceptions (Berry and Linoff, 1997). Perceptions are the first artificial neural networks, which are based on a unit called the perception, which produces in output scaled as 1 or -1 depending upon the weighted, linear combination of inputs' (Fraser, 2000).

Later, in the 1960s two scientists demonstrated basic theoretical deficiencies of the above networks (Berry and Linoff 1997). The scientists (Minsky and Papert) demonstrated that the perception could not represent simple functions, which were linearly inseparable (Stergiou and Siganos, 1996).

The above deficiency led to the decline in the study of neural networks in the 1970s (Fraser, 2000). Then, in 1982, John Hopfield invented back propagation, a way of training neural networks that sidestepped the theoretical pitfalls of earlier, approaches.

This resulted in the renaissance of neural network research and in the 1980s neural network researchers moved from the lab to the commercial world where they have since been applied in virtually every industry (Berry and Linoff 1997).

Several factors contributed for the popularity of neural networks in the 1980s. First, computing power was available in abundance. Second, analysis became more comfortable with neural networks realizing that they are closely tied with familiar statistical methods. Third, data was available easily because of automation of most operations (Berry and Linoff 1997).

2.3.1.1.2 Application

These networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. According to Berry and Linoff (1997), Bigus (1996) and Atiezenza et. al. (2002) Neural Networks are probably the most common data mining techniques, perhaps synonymous with data mining to some readers.

Rogers and Joyner (2002) described neural networks as a computer application that mimics the neuro-physiology of the human brain, which is capable of learning from examples to find patterns in data. Neural networks are associative self-learning techniques with the ability to identify multidimensional relationships and perform pattern recognition in non-linear domains (Atienza et. al., 2002).

Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions (Bigus, 1996). Neural networks may be used in classification problems (where the output is a categorical variable) or for regression (where the output variable is continuous) (To crows corporation, 1999).

Up to date, neural networks have already been put to use to solve different problem domains. Some of these domains of problems as described by Frohlich (1999), Berry and Linoff (1997) are:

- Pattern association
- Pattern classification
- Regularity detection
- Prediction
- Clustering
- Optimization problems etc.

Some of the specific application areas where neural networks have been put to use include (Stergiou and Siganos 1996-97; Lawrence, 1994, Fraster, 2000 and Smith, 1998):

- Credit risk assessment.
- Neural networks for handwritten character, speech, fingerprint and electrical signal recognition.
- Neural networks that detect hypertension and heart abnormalities.
- Neural network for automatic vehicle control.
- Neural networks that are used in improving marketing mail shots etc.

2.3.1.1.3 Basic Structure

The structure of neural network is very similar to the structure of the neurons in our brain. The neuron consists of ‘dendrites for incoming information and axon with dendrites for outgoing information that is passed to connected neurons’ (Frohlich, 1999). Information is transported between neurons in the form of electrical stimulation along the dendrites. The incoming information that reaches the dendrites are added up and passed along the axon to the dendrites at the other end. At the synapse transmission of signals from one neuron to the other takes place.

Grove (1996) discusses that the brain is composed of around 10^{11} neurons, which are arranged in a rough layer like structure. Each neuron works as a processor and massive interaction between all cells and their parallel processing makes the brain’s abilities possible (Frohlich, 1999). The early layers receive information from the sense organs (eyes, ears). The final layers produce motor output (e.g. Moving arms and legs). The middle layer forms the associative layer. This layer is the least understood but is considered to be the most important part of the brain in humans.

Frohlich (1999) write that neurons are adoptive i.e. the connection structure is changing all the time and learning takes place through these adoptions. Grove (1996) also put that the brain seems to learn in three ways by growing new axons, by removing axons and by changing the strengths of existing axons.

The simulation of the human brain is what we call artificial neural network or simply neural network (Smith, 1998). ‘However, because our knowledge of neurons is incomplete and our computing power is limited, our models are necessarily gross idealizations of real networks of neurons’ (Stergiou and Siganos, 1996).

Information is received from different inputs and are combined using a combination function, which is usually a summation of the different weights. Then a transfer function is used to calculate a single output between 0 and 1. The transfer function represents the non-linear characteristics exhibited in biological neurons.

Typical functions include threshold function, step transfer function, sigmoid function and Gaussian functions (Lawrence, 1994). Among the different functions the common one is the sigmoid transfer function. (Fraser, 2000; Frohlich, 1999). This is also the function used for this research work since the provider's of the software employed, put that they had not seen any problem that train fundamentally better with anything other than the standard sigmoid transfer function (California Scientific Software, 1998). A mathematical description on the different functions is available in the work of Lawrence (1994).

The formula for sigmoid function is given by:

$$v = (1 + e^{-s})^{-1}$$

Where s = sum of the inputs to the neuron

v = value of the neuron

The combination and transfer function together make up the activation function (Fraser, 2000). The resulting value from the activation function will be compared with certain threshold value. If the input exceeds the threshold value, the neuron will be activated, otherwise it will be inhibited. If activated, the neuron sends an output on its outgoing weights to all connected neurons and so on (Frohlich, 1999).

In artificial neural networks, neurons are grouped in layers (Frohlich, 1999). There are basically three types of layers. These are the input, hidden and output layers which constitute of the input, hidden and output neurons respectively

(Frohlich, 1999). Knowledge Technology Inc. defines the three terms as in below.

Input Layer: A layer of processing elements that receives the input to the neural net.

Hidden Layer: A layer of processing elements between a neural network's input layer and its output layer.

Output Layer: The layer of processing elements, which produce a neural net's output.

There could be a number of input, hidden and output neurons in their respective layers. For instance in the following neural network, there are four input, four hidden and three output neurons. And the network has one input, one hidden and one output layer.

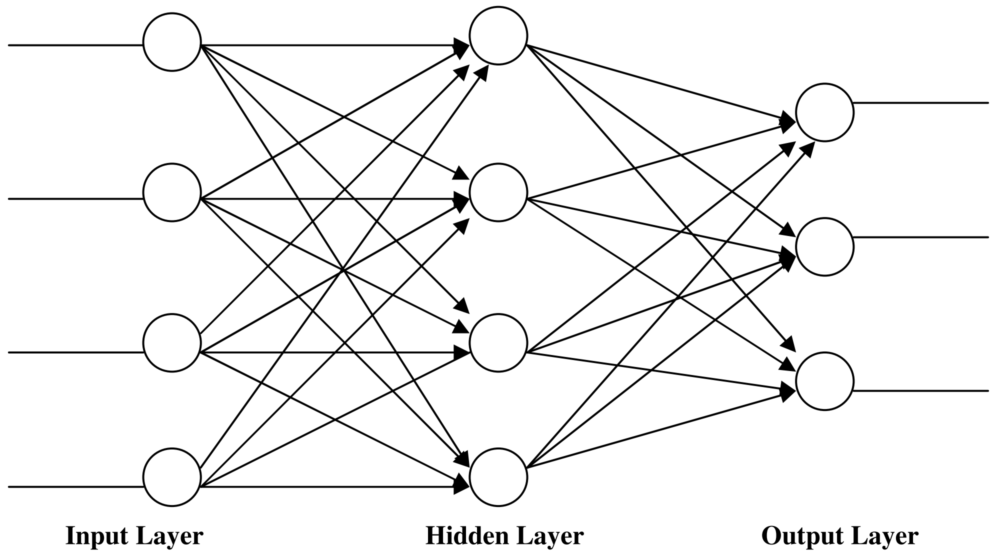


Figure 2: A simple Feed Forward Neural Network

The way neurons are connected and the learning rule adopted determines the learning process in a neural network. These concepts are discussed in the next section.

2.3.1.1.4 Classification of Neural Networks

Lawrence (1994) describes that neural networks can be described ‘in terms of the connections between them (topology) and its learning rule.’

Connections (topologies) define how data flows between the input, hidden and output layers, and these usually have an enormous effect on the operation of a network (Lawrence, 1994).

According to Bigus (1996) there are two major categories of connection topologies. These are, feed forward and recurrent networks also known as feedback networks. In feed forward network, information flows from input layers to zero, one or more succeeding hidden layers and then to the output layer i.e. data travel only in one way - there are no feedback or loops. First, data enters the neural network through the input neurons. The total input signal is then passed through an activation function in order to determine the output.

Feed forward networks are used in situations where all the information to bear on a problem can be presented at once. For instance, for the research being undertaken the feed forward networks would be employed since a network is to be trained by providing information related to a child together with the outcome i.e. whether that child with the stated information is dead or alive.

The other types of networks are the recurrent networks (feedback networks). In feed back networks signals can travel in both directions by introducing loop in the network (Stergiou and Siganos, 1996). In recurrent networks, information about past inputs is fed back into and mixed with the inputs through recurrent or feedback connections for hidden or output units. This makes the neural network to contain memory of past inputs. Such networks are used in situations where we need the neural network to somehow store a record of prior inputs and factor them in with the current data to produce an output.

Neural Networks could also be discussed in terms of their learning mechanisms. The learning rule is the very heart of a neural network. It determines how the weights are adjusted at the neural network gains experience (Lawrence, 1994). Many networks use some variation of the Delta rule for training. One type of rule most widely used is back propagation (Frohlich, 1999; Fraser, 2000 and Knowledge Technology, Inc., 2000).

The back propagation algorithm is also the algorithm considered in this study since it is best suited for a classification activity (Bigus, 1996). The back propagation algorithm uses the supervised learning approach during training.

In supervised learning, the network learns through examples. In particular, the neural network is given a problem and it makes a classification or prediction. Then at that point the network would be given the correct answer. The learning algorithm takes the difference between the correct output and the prediction the neural network made and then uses the information to adjust the weights of the neural networks so that the prediction next time would be closer to the correct answer. The neural network has to be given examples many times in order to learn and make correct predictions (Bigus, 1996).

The other learning paradigm is the unsupervised learning where a teacher is not required for training (Grove, 1996). In this paradigm, there are no target outputs and it is impossible to determine what the result of the learning process will look like (Frohlich, 1999). The network is simply exposed to number of inputs and the network organizes itself in such a way as to come up with its own classification for inputs (Lawrence, 1994)

As mentioned, the back propagation uses the supervised training paradigm. A typical back-propagation algorithm consists of three steps. First, the input pattern is presented to the network whereby the input patter is propagated through the network until they reach the output units. This forward pass would produce the actual or predicted output. Then the desired output would be given as part of the training set, so that the actual output can be subtracted from the desired output in order to give the error signal. In the third step the errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output. The connection weights are then adjusted and the neural network is said to have learned from an experience (Bigus, 1996). These three steps are repeatedly carried out for every example in the data until the weights are no more adjusted.

2.3.1.2 Decision Trees

Another most commonly used data mining techniques for classification tasks are decision trees. Decision trees are simple knowledge representation and they classify examples to a finite number of classes. In decision tree induction, the nodes of the tree are labeled with attribute names, the edges of the tree are labeled with possible values for the attributes and the leaves of the tree generate decision tree from a given set of attribute-value tuples. A decision tree is constructed by repeatedly causing a tree construction algorithm in each generated node of the tree (Larvac, 1998). The classification is performed separately for each leaf, which represents a conjunction of attribute valued in a rule (Last and Kandel, 2002).

Structurally, decision trees consist of two types of nodes; non-terminal (intermediate) and terminal (leaf). The former correspond to questions asked about the characteristic features of the diagnosed case. Terminal nodes, on the other hand generate a decision (Rudolfer, Paliouras, and Peers, 2002).

Decision trees are a way of representing a series of rules that lead to a class or value. By navigating the decision tree it is possible to assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached (Two crows corporation, 1999).

Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), Quest, and C5.0 (Two crows corporation, 1999).

An important quality of decision trees is that they handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets (Two crows corporation, 1999). Rogers and Joyner (2002) also stated "tree-based models are good at selecting important variables, and therefore work well when many of the predictors are irrelevant."

A common criticism of decision trees is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In addition, all splits are made sequentially, so each split is dependent on its predecessor (Two crows corporation, 1999).

2.3.2 Regression Models

The prediction of continuous values can be modeled by statistical techniques of regression. Many problems can be solved by linear regression, and even more can be tracked by applying transformations to the variables so that a nonlinear problem can be converted to a linear one (Han and Kamber, 2001). Levin and

Zahavi (1999) stated that regression models are the leading predictive models. The most common regression models are, linear regression models (for modeling continuous response), and logistic regression models (for modeling discrete choice response) (Levin and Zahavi, 1999). These are regression models that take into account the effect of time. The most common types of time dependent models are: survival models, time-series models, non-stationary models in which the course of the event spell, time related censored and truncated data, and others (Levin and Zahavi, 1999).

2.3.3 Clustering Models

The process of grouping a set of physical or abstract objects into class of similar objects is called clustering (Han and Kamber, 2001). A cluster is a set of objects grouped together because of their similarity or proximity. Objects are often decomposed into an exhaustive and/or mutually exhaustive set of clusters (Rea, 2001). It is a process of mapping a data item into one of several clusters, which are not pre-specified but are determined from the data (Levin and Zahavi, 1999). Clusters are formed by finding natural groupings of data items based on similarity matrices, proximity considerations and probability measures. By clustering, one can identify dense and spare regions and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research (Han and Kamebr, 2001).

In business, clustering can help marketers discover distinct groups in their cluster bases and characterize customer groups based on purchasing patterns. Clustering may also help in the identification of areas of similar land use in an earth observation database, and in the identification of groups of automobile insurance policy holders with a high average claim cost, as well as the identification of groups of horses in a city according to house type, value, and geographical location. It can also be used to help classify documents on the web for information discovery (Han and Kamber, 2001).

There are a number of approaches for mining clusters. One approach is to form rules, which dictate memberships in the same group based on the level of similarity between members. Another approach is to build set functions of some parameters of the partition (Rea, 2001).

Levin and Zahavi (1999) argued that perhaps the most common of all automatic clustering algorithms is the K-means algorithm, which assigns observations to one of K classes to minimize the within-cluster-sums-of-squares. Also worth mentioning are the Judgmental-based or "manual" segmentation methods which are still very popular in direct marketing applications to carve up a customers list into homogeneous segments (Levin and Zahavi, 1999).

2.3.4 Link Analysis

Link analysis is concerned with finding rules between data elements. The two most common rules are association and sequencing rules (Levin and Zahavi, 1999).

Association methods discover rules of the form: "if item *A* is part of an event, then *X%* of the time item *B* is also part of the event" (Levin and Zahavi, 1999). Given a collection of items and a set of records, each of which contain some number of items form the given collection, an association function is an operation against this set of records which return affinities or patterns that exist among the collection of items (Rea, 2001).

To assess association of items in the database, two probability measures, called support and confidence, are introduced. The support (or prevalence) of a rule is the proportion of observations that contain the item set of the rule. The confidence is the conditional probability of B given A, P (B/A). Levin and Zahavi (1999) argued that a rule is “interesting” if the conditional probability P(B/A) is significantly different than P(B). A typical application, identified by IBM that can be built using an association function is Market Basket analysis (Rea, 2001). This is where a retailer runs an association operator over the point of sales transaction log, which contains among other information, transaction identifiers. The set of products identifiers listed under the same transaction identifier constitutes a record. The output of the association function is, in this case, a list of product affinities (Rea, 2001).

Another example of the use of associations is the analysis of the claim forms submitted by patients to a medical insurance company. By defining the set of items to be the collection of all medical procedures that can be preformed on a patient and the records to correspond to each claim form, the application can find, using the association function, relationships among medical procedures that are often performed together (Rea, 2001).

2.4 Data Mining Applications in the Health Care

Widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer-assisted analysis (Larvac, 1998). Extensive amounts of data gathered in heath care databases require specialized tools for storing and accessing data, for data analysis, and for effective use of data. Medical informatics may use the technologies developed in the new interdisciplinary field of knowledge discovery in databases (KDD) (Larvac, 1998), and particularly data mining (Lloyd-williams, 2002). Today, numerous health care organizations are using data mining tools and techniques in order to raise the quality and efficiency of health-related products and services. A number of articles published in the health care literature revealed the practical application of data mining techniques in the analysis of health information.

The reminder of this section attempts to review some empirical studies related to the practical application of data mining technology in the health care sector.

In an effort to turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficiency of patient care. Data mining can be used to help predict future patient behavior and to improve treatment programs. By identifying high-risk patients, clinicians can better manage the care of patients today so they do not become the problems of tomorrow (Rogers and Joyner, 2002).

Current trends in medical decision making show awareness of the need to introduce formal reasoning, as well as intelligent data analysis techniques in the extraction of knowledge, regularities, trends and representative cases from patient data stored in medical records (Larvac, 1998). According to Larvac, machine-learning methods have been applied to a variety of medical domains in order to improve medical decision-making.

Prather, et. al. (2001) conducted a data mining project at Duck University Medical Center using an extensive clinical database of obstetrical patients to identify factors that contribute to prenatal outcomes. The goal of this knowledge discovery effort was to identify factors that will improve the quality and cost effectiveness of prenatal care. The production system database identified for mining was the computer-based patient record system known as "The Medical Record", or TMR. TMR is a comprehensive longitudinal clinical patient record system (CPRS) developed at Duke University over the last 25 years (Prather, et. al., 2001). For their work, the specific database selected for the data mining project was the prenatal database used by the Department of Obstetrics and Gynecology at Duke University Medical Center. For the purpose of the initial study, the researchers created a sample two-year dataset (1993-1994) from the data warehouse. Exploratory factor analysis was selected for data mining. The statistical software used to conduct the factor analysis was SPSS for windows version 5.0 (Prather, et. al., 2001).

Prather and his co-workers confirmed that a large clinical database could be successfully warehoused and mined to identify clinical factors associated with preterm birth. Finally, the authors concluded that data warehousing and mining technology are applicable to health care, and that the preliminary mining of a clinical data warehouse has produced promising results.

Last, Maimon and Kandel (2002) have applied the process of knowledge discovery to a dataset of 33,134 mortality records extracted from the Israeli Ministry of Health mortality database. The data set analyzed by these researchers includes the records of all Israeli citizens who passed away in the year 1993. The purpose of this study was to identify the leading causes of death and the association of various factors with certain diseases by applying data mining techniques. The death cause (medical diagnosis) of each person was defined by the International 6-digit code (ICD-9-CM). To carryout the project successfully, a significant amount of data preparation was performed. The information-theoretic approach to data mining was the approach that has been used by these researchers. According to the authors, the data mining process has resulted in selection and scoring of the most important input attributes, discretization of continuous features, rule extraction, and calculation of data reliability. Finally, the following main results were obtained:

- The automated data cleaning procedure has revealed outliners in most attributes of the database.

- Traditionally, the mortality data has been analyzed with respect to age, gender, and ethnic origin. However, the results achieved by their study suggested that time of year are more important factor for death than place of birth and/or ethnic origin.

- Rules defining high-risk and low-risk groups (w.r.t specific causes) have been extracted and scored by the information theoretic network. The researchers argue, these rules can be used for determining priorities in the health care budget. They may be also valuable for insurance companies and other commercial institutions.

- Most unreliable records in the database contain lowly probable information. The authors further proposed that this information should be checked by medical experts and possibly compared to the manual source. This comparison can lead to correcting the data in the original database.

Lloyed-Williams (2002) had also analyzed datasets extracted from the World Health Organization's Health for All (HFA) Database using a data mining approach. During the selection process, mortality data based on the following conditions was extracted by the researchers from the HFA database: life expectancy at birth; probability of dying before five years of age; infant mortality; post-neonatal mortality; standardized death rate (SDR) for circulatory diseases; SDR for malignant neoplasm; SDR for external causes of injury and poisoning; SDR for suicide and self-inflicted injury. Data was extracted for 39 European Countries, and then converted into a format acceptable to the software used for that particular project.

An underlying aim of the study, as noted by the author, was to track changes in the data that may have occurred over the years for the same samples of countries in order to examine whether any patterns identified remained consistent over time. The extracted data was analyzed by custom written Kohonen self-organizing map software in order to identify possible groupings. Standard statistical techniques were used to evaluate the validity of the groupings (Lloyd-Williams, 2002).

Preliminary work of Lloyd-William's study resulted into two groups for clusters of countries in each year being apparent. In addition to the geographical division, the classification also appeared to reflect differences in wealth. Countries in the first of the groups were relatively poor; where as countries in the second of the groups were relatively wealthy. Lloyd-Williams reported that the observation that the classification appeared to reflect two different GNP groups suggested that GNP could be inter-related with the health indicators. In order to further explore this possibility, the author calculated coefficient of correlation between GNP and all the seven HFA indicators he has used in the initial analysis. Results obtained indicated that GNP is strongly and positively correlated with life expectancy, and strongly but negatively correlated with the SDR for diseases of the circulatory system.

Downs and Wallace (2001) have also applied data mining techniques to mine association rules from a pediatric primary care decision support system. According to the authors, the purpose of their study was to apply an unsupervised data-mining algorithm to a database containing data collected at the point of care for clinical decision support. They took the data set from the Child Health Improvement Program (CHIP), a preventive services tracking and reminder system in use at the University of North Carolina. The workers used the unsupervised data-mining (pattern discovery) algorithm to extract 2nd and 3rd order association rules from the data. As a result of the data mining process, the algorithm, which the authors have used, discovered 16 2nd order associations and 103 3rd order associations. The authors have also identified that the 3rd order associations contained no new information. However, the 2nd order associations demonstrated a covariance among a range of health-risk behaviors. Additionally, the algorithm discovered that both tobacco smoke exposure and chronic cardiopulmonary disease are associated with failure on developmental screens.

Summarizing their results, Downs and Wallace (2001) stated that the discovery of a direct association between cardiopulmonary disease (e.g., asthma) and developmental delay among otherwise healthy children was a novel discovery. However, the literature shows a high covariance among a range of health risks that may explain the coexistence of these problems in impoverished families.

Resnic et.al. (2002) studied 2,804 cases from January 1997 through February 1999, in order to develop risk models to predict death, and post-procedural myocardial infraction following precutaneous coronary intervention (PCI). The researchers constructed risk models using multivariate logistic regression, artificial neural networks and prognostic risk scoring systems. As a result of their investigation, the researchers reported that, composite logistic regression models and artificial neural networks performed similarly in predicting the risk of major acute complications (i.e. predicting 0.812 and 0.807, respectively). According to the researchers, the risk scoring models appeared to provide reasonable discrimination while offering the potential for simple clinical implementation in the estimation of the risk of death and myocardial infraction in interventional cardiology.

Timm (1998) evaluated data mining procedures and its results using a data set sampled in cooperation with the anesthesia ICU of the central hospital St.-Juergen-strabs Bremen. Prediction of medically relevant risks like infection or length of stay in the ICU have been defined as major tasks for the data mining process. The researcher has also compared the results of classical risk scores with generated risk classifications of mortality. The data mining process was applied to a database consisting of all variables maintained on a daily base for 400 patients.

For hypothesis generation, the researcher used methods from machine learning, statistics and neuro-science and he compared their results. As a result, the researcher reported that all the approaches proved to yield comparatively sufficient results for practical use as far as misclassification rates come into consideration. On the other hand, they presented great differences with respect to specificity and sensitivity. Only the neural network classification achieved sufficient results concerning sensitivity (95%) and specificity (93%)(Timm, 2002).

Wang et. al. (2001), have also built several logistic regression and neural network predictive models to estimate the likelihood of myocardial infraction based upon patient reportable clinical history factors only. They used two datasets totaling 1753 patients to build and test their models. The first dataset consisting of 1253 patients was used to build and train the models. These researchers reported that both logistic regression and neural network models were constructed using all variables, and using standard stepwise, forward, and backward variable selection algorithms. The most commonly selected variables were then used to construct the best model. Then, the best performing logistic regression and neural network models were chosen and evaluated using previously unseen data set of 500 patients. The final logistic regression model had a c-index of 0.8444 on their validation data set, and the neural network showed a c-index of 0.8503 on the same validation set.

As a result of their experiment, Wang et. al. (2001) concluded that logistic regression and neural network models can be built that successfully predict the probability of myocardial infraction based on patient-reportable history factors alone. So, models that only require patient reportable factors for prediction may have important applicability as screening tools in settings outside of the hospital when patients need advice on whether or not to seek professional care (Wang et. al., 2001).

Plate et. al. (1997) reported on a comparative study of the predictive quality of neural networks and other flexible models applied to real and artificial epidemiological data. The authors made a comparison between neural networks and other statistical techniques for modeling the relationships between tobacco and alcohol and cancer. They used a data set consisting of information on 8,562 subjects from an epidemiological study conducted by the division of epidemiology and cancer prevention at the BC cancer agency. These researchers compared models based on their performance on the held-out test data, so as to avoid over fitting bias in evaluation.

As a result of their experiment, the authors concluded that for predicting the risk of cancer in their data, neural networks with Bayesian estimation of regularization parameters to control overfitting performed consistently but slightly better than logistic regression models. They further stated that given their ability to not overfit while still identifying complex relationships, neural networks could prove useful in epidemiological data-analysis by providing a method for checking that a simple statistical model is not, missing important complex relationships.

Theeuwen, Kappen and Neijt (2001) have also analyzed the difference between the traditional Cox regression analysis and artificial neural networks to model survival analysis to predict treatment outcome in patients with ovarian cancer.

The researchers constructed a database including 917 patients from 4 studies, two studies from the Netherlands joint study group for Ovarian Cancer and two from the gynecological cancer cooperative group of the Europe an Organization for Research and treatment of cancer (EORTC). For each year they trained a Boltzmann perceptron to classify the patients into the classes Alive and Died, thereby they model the survival probability for each year.

According to these authors, in order to obtain a minimal set of prognostic factors with a maximal predictive recall, the set of prognostic factors were carefully reduced. To do this, all patient characteristics were normalized such that the magnitudes of their weights could be compared. Then, the researchers trained the neural network until convergence was obtained, after which the patient characteristic with the smallest attribute weight was removed. The researchers reported that this procedure was repeated until the performance on the independent test set was deteriorated.

Finally, the authors concluded, a pilot comparison between Cox's survival analysis method and the Boltzmann method showed that the neural network performed slightly better. As a justification to this output, the authors argued, this can mainly be attributed to the paradigm of training and test set, which leads to a better prediction of independent data than standard statistical tests assuming Gaussian distribution of errors.

CHAPTER THREE

THE DESIGN AND ANALYSIS OF EPIDEMIOLOGICAL STUDIES

3.1 Overview of Epidemiological Studies

The health of populations is a major concern. Health improvement programs and health actions is a major initiative to improve the health of the population. Epidemiology is a major tool for understanding the incidence and spread of disease in the population and therefore provides essential information for various health related programs (Moon et. al., 2000).

Epidemiology, the science of the distribution and determinants of health-related states in human population, is a basic methodological science for medical research with inputs from several disciplines, including those outside the traditional medical field (Last, 1988). The concepts of "distribution" and "determinants" are important elements in the study of disease. The former implies that interest is focused on describing how the health status of the population relates to socio-demographic characteristics, time periods and a host of other variables. The latter implies searching for causes of disease patterns (Desta, 1994).

An important principle of epidemiology is that human disease does not occur randomly. Diseases may have different distributions depending upon the underlying characteristics of the populations being studied (Moon et. al., 2000).

Disease determinants are factors or events that are capable of bringing about a change in health. Some examples are specific biological agents that are associated with infectious disease or chemical agents that may act as carcinogens. For the epidemiologist, a key question is the extent to which a group membership is linked to a disease determinant. Attention focuses on asking why certain diseases concentrate among particular population groups. So, another important principle of epidemiology is that human disease has causal and preventive factors that can be identified via systematic investigation of who gets ill (Moon et. al., 2000).

According to Frelichs (1991), epidemiological surveillance is an on-going system for the collection, analysis and interpretation of health data essential to the planning, implementation, and evaluation of public health programs, closely integrated with timely dissemination of these data to those who need to know. He added that the final link in the surveillance chain is the application of these data to prevention and disease control. As Desta (1994) puts it, epidemiological surveillance activities are usually included as part of the health information system of governments that have assumed responsibility for health care. Field personnel are expected to complete reporting forms for the health information system on a regular basis and send the forms to the immediately superior health care unit.

However, there has to be a balance between the time spent on data collection and services. If too much time is spent on collection of data alone the efficiency with which the program is managed may decrease. Surveillance by definition implies action orientation. These data should be transformed into information that can easily be communicated to the managers for decision-making (Desta, 1994).

Desta (1994) further stated that in developing countries, surveillance data are often presented in complicated formats that are difficult to comprehend and which do not point to any specific action. As the manager is often tempted to make decisions without data, this will leave him a further excuse for not working about making informed decisions, and he will then allocate resources in haphazard fashion. Unfortunately, decisions on new developments or changes in health services are often initiated on emotional or political grounds without careful consideration of the consequences. Luck of adequate health information systems has been identified as one of the constraints in health service delivery in Africa (Yemane et. al., 1999).

Why Epidemiological Studies?

Appropriate planning, management, and evaluation of health programs depend to a large extent on the access to timely and accurate data concerning demographic characteristics, on the occurrence of major health problems, and on associations with underlying factors. Population based data are prerequisites for such situations, and form the basis for valid health information. Epidemiological principles and methods are an aid to the development of such systems (Desta, 1994).

The epidemiological approach plays a key role in the search for disease determinants and for the planning and evaluation of community interventions and health services. This population perspective is of particular relevance in developing countries. However, Third World epidemiology suffers from several conceptual and structural syndromes. First, health problems are so obvious that research may seems as just postponing necessary actions, Epidemiology is however mandatory both for the planning and for the evaluation of such interventions and should challenge a common misconception among health planners and experts that the rural population in many developing countries live under homogenous, equally bad, conditions. Second, epidemiology in developing countries has mostly taken a descriptive mode, generating numerous surveys but less analytical knowledge. The lack of statistical sampling frames has “Justified” the haphazard and accidental selection of study objects, high migration figures have “motivated” for many cross-sectional studies at the expense of more analytical designs. Also the export of data for analysis at Western research institutes has been “sanctioned” by measure data processing facilities. The general lack of resources, therefore, is a motivation for modern epidemiology with appropriate microcomputer technology and for cost-effectiveness in study design (Desta, 1994).

Organizing Epidemiological Studies in Developing Countries

Epidemiological studies in developing countries have often been cross-sectional in design. Appropriate sampling frames and population registers are often difficult to achieve in view of the high migration figures (Desta, 1994). To solve this problem, well-designed epidemiological studies in defined populations are needed to provide a representative picture of the health status of the entire population. However, as Ekanem (1985) puts it there are a number of difficulties encountered in carrying out such studies in developing countries. Lack of basic demographic data covering the total population, absence of vital events registration systems, and other related issues are the major constraints faced in trying to carry out population based surveys. According to Desta (1994), this has forced researchers to use institution-based data sources such as health care units, and to use cross-sectional designs. Although institution based data are useful in their settings, they are of limited value in describing the health status of the total population.

The other important issue to be considered in establishing epidemiological studies in developing countries is the method to be used to gather population-based data. As Desta (1994) wrote, in countries where there is a high rate of illiteracy, self-administered questionnaires become an impossible means of data collection. So, one has to rely on interviewers, who have to be recruited and given appropriate training. The degree of cooperation one gets from the study subjects may also be influenced by the level of literacy.

Generally, while epidemiological research in developed countries under ideal circumstances is facilitated by the availability of demographic data, population and disease registers, efficient and well established vital statistics systems, good communication systems, enlightened societies, government support through generous funding and the availability of advanced hard-and software, the extreme opposite exists in many developing countries (Desta, 1994). He further argued that conducting research and establishing epidemiological studies in developing countries often requires monumental dedication and single-mindedness in the face of the day-to-day difficulties rarely faced by researchers in developed countries.

3.2 The Butajira Rural Health project (BRHP)

In Ethiopia, as in other developing countries, adequate and reliable health information is lacking (Desta, 1992). The capacity in Ethiopia for collecting, compiling, analyzing, interpreting, and disseminating the appropriate information for decision making is very poor. There is no systematically organized registration of vital events. Population-based studies are rarely carried out in Ethiopia. Moreover, national figures are of limited use for health planning at regional levels (Desta, 1992).

The Ministry of Health of the Federal Democratic Republic of Ethiopia bears the primary responsibility for the health care of the Ethiopian citizens. The Ministry is also responsible for the constant improvement of health services, including the support of medical research activities. To achieve its purposes, the ministry has adopted the PHC (Primary Health Care) strategy, which includes expansion and updating of health services with emphasis on disease prevention particularly in undeserved rural populations (Desta, 1994). Such a strategy implies that data on major health problems of defined population groups and on coverage and usage of health services are essential in the planning and evaluation of health services.

As Desta (1994) stated, present information, however, is mainly based on records of patients of health care units. Community based studies, which relate the provision and utilization of health services to people's needs and demands are lacking.

Indeed, community diagnosis by identifying major health problems of the community requires basic information on demographic characteristics of the population. The availability of valid data on demographic and epidemiological conditions, and on patterns of health care units and their utilization, facilitates planning, monitoring and evaluation of health services (Desta, 1992). The author further stated that longitudinal population-based studies are needed to generate sound data on morbidity, mortality, and fertility.

The Butajira Rural Health project in Ethiopia was the first population-based epidemiological study established in 1986 to support research and intervention programs in primary Health care. The ultimate aim of the project was therefore to provide current epidemiological information system and thereby to contribute to the improvement of primary health care management and decision-making (Desta, 1994). Since the basic reason for this type of population-based surveillance arises out of lack of civil registration infrastructure, it follows that appropriate methodologies have to be implemented which are based on epidemiological methods and principles. Thus in many ways the BRHP has pioneered methods of epidemiological and demographic surveillance that can be appropriately applied in this kind of setting (Yemane, et.al., 1999).

3.2.1 The Project Design

Yemane et. al. (1999) wrote that the objectives of the BRHP, when initiated in 1986, were to generate health-related information, to establish and maintain an epidemiological research laboratory, to build and strengthen research capability and to develop local capacity in the prevention and control of disease.

The study district is administratively located in Gurage Zone, Southern Nations, Nationalities and peoples Regional State (SNNPRG). The district is divided for administrative purposes into a total of 82 peasants' Associations (PAs) in the rural areas and 4 Urban Dwellers Associations (UDAs) in Butajira town (Yemane et. al., 1999). According to Yemane et. al. (1999), the project has been conducted in a set of nine randomly selected PAs (using the probability proportionate to size technique) and one UDA. In line with its objectives, the initial tasks for the BRHP during 1986-87 were to perform a census of the population in the selected villages to obtain the baseline population and to establish a system of demographic surveillance with continuous registration of vital and migratory events at a household level. Events registered by the BRHP are birth, death, cause of death, marriage, new household, out-migration, in-migration and internal move (migration within the BRHP surveillance villages).

3.2.2 Population Development (1987-96)

Since January 1987 there has been continuous surveillance of the 10 selected PAs in Butajira district, with re-census process in 1995 and 1996. The total population of the BRHP enumerated during the first census conducted at the beginning of 1987 was about 28, 616 individuals. The data that was collected initially included basic demographic characteristics (like age, sex, etc.), cultural patterns, household structure, environmental conditions, health care utilization, and recall of vital events.

The aim of the census was to enumerate the study population (i.e. to get a baseline for the continuous surveillance) and to get some background information about the population. After the census was completed, demographic surveillance with monthly visits to individual houses was introduced and still on going. Information collected included changes that had occurred since the previous visit - births, deaths, causes of death and migration.

On scientific grounds, periodic census is necessary to check on the validity of the surveillance system and to detect any missed events. To achieve this purpose, a second census was done in November 1995 in all the ten study villages. Since the interval between the two censuses was felt to be too long, a third census took place in November 1999. The result of these censuses showed an extent of similarity between the census and surveillance database, which illustrates the quality of the continuous registration system (Yemane et. al., 1999).

As Yemane et. al. (1999) stated, during the ten-year period (1987-1996), the study population was observed to increase from a baseline population of 28, 616 to 37, 323 at the beginning of 1997, implying a mean yearly growth rate of 2.7%. During this time a total of 5,143 deaths and 15,667 births were registered in the area. From overall mortality, under-five mortality accounted for 50% of all deaths.

3.2.3 Data Collection Methods[1]

At the BRHP, surveillance data is collected monthly by visiting each household, and each household is identified by a unique number within its village, and each individual within their household.

Individuals can thus be characterized by a unique identity number with in the entire surveillance system. This unique identity number is ten characters long and is composed of the village number, the household number and the individual number within the household. Once an individual is allocated a unique identity number in this way, it is permanently associated with that individual, even if they move to a new location (the detail, of which are separately recorded).

[1] Yemane Berhane, et al., 1999. Establishing an epidemiological field laboratory in rural areas- potentials for public health research and interventions. *The Ethiopian Journal of Health Development,* vol.134, special issue.

Any adult of the household above the age of 15 years is eligible to respond to the monthly household event collection interviews. The state of each individual is checked during the monthly visit and recorded on a demographic surveillance sheet. Basic demographic, social, housing conditions and health care utilization characteristics are recorded for each household at entry into the surveillance system and during any re-census process.[1]

The surveillance system operates on an open cohort system and is dynamic. Individuals enter and leave the system any time. Since individual movements are tracked regularly, an individual's overall time contribution to the study base can easily be calculated. The individual person-times calculated can then provide denominators for calculation of, for example, mortality and fertility rates within the study base.

Each vital event is registered on a separate form at the household level. Each form contains several questions that are pertinent to adequately characterize the event. Household interviews are conducted by using village-based enumerators. Village-based enumerators have had at least 10 years of formal education. They were trained at recruitment on the use of the surveillance forms, on how to conduct interviews and on data management in the field. In addition, they receive periodic refresher training on the same subjects. Data collection is usually undertaken in the morning (0700-1300). Each village is divided into four zones for the sake of simplicity. The enumerator in the village will spend one week in each zone and is not allowed to continue onto the next zone before the allotted time even on completing one zone. This is to ensure that each household is visited at monthly intervals. Then, completed questionnaires are submitted by the enumerators to their field supervisors on a weekly basis.

[1] Yemane Berhane, et al., 1999. Establishing an epidemiological field laboratory in rural areas- potentials for public health research and interventions. *The Ethiopian Journal of Health Development,* vol.134, special issue.

3.2.4 Data Quality Assurance[1]

Data quality assurance mechanisms have been instituted at several points. The most critical of these is the field supervision. Field supervisors (4 persons) each designated to 2-3 villages and a project coordinator at Butajira perform the immediate supervision of data procedures on daily basis. These people have served the project from its inception and are vital organs of the surveillance system. Their tasks include checking of each and every completed surveillance form and visiting a randomly selected 5% of households each month on a weekly-distributed timetable. The next level of supervision is performed by the BRHP research assistants (1-2 persons). Research assistants have had public health training at Masters' level. They are responsible for the overall supervision of the surveillance system. They also perform actual data checking at a field level and randomly check on some households every fifteen days.

3.2.5 Data Entry to the Main Database[1]

Data entry is performed at the Department of community health in Addis Ababa. Specifically developed software, using the dBase IV platform, is used to facilitate automatic data checking at entry. Each event form is entered in a separate transaction file by data entry clerics. The transfer of data from the transaction files to the main database is done by the researchers using a monthly updating function. Data records rejected by the computer for various reasons are returned to the field supervisors for appropriate measures through the research assistant. The introduction of the above computer software has substantially improved data quality and management.

[1] *Yemane Berhane, et al., 1999. Establishing an epidemiological field laboratory in rural areas-potentials for public health research and interventions. The Ethiopian Journal of Health Development, vol.134, special issue.*

3.2.5.1 Structure of the BRHP Database System[✏✏]

The software currently being used for the BRHP data management system was written using dBase IV, running under DOS. The program is in a file called MAIN.PRG, which resides with other files in a folder called BUTA. The software has been deliberately designed for the Ethiopian setting, particularly in respect of the dual calendar system in use (the Ethiopian calendar for field work, and the international calendar for later analysis). The software provides a user interface that imposes checks and restrictions on how individual records are handled. It also provides a mechanism for receiving data from the field, checking it, and raising queries as needed when data are incomplete or inconsistent.

Within the system, there are two important files that contain the current state of the population under surveillance. These are a file of individual census records, and another relating to the households where the individuals live. When event such as births and deaths occur in the population, they are noted by the enumerators and entered into appropriate sub-files. These are known as transaction registers, and can then by used to update the main individual and household files in a secure manner. This structure is presented diagrammatically in the following diagram.

[✏✏] ***A Manual of the BRHP database: Unpublished material***

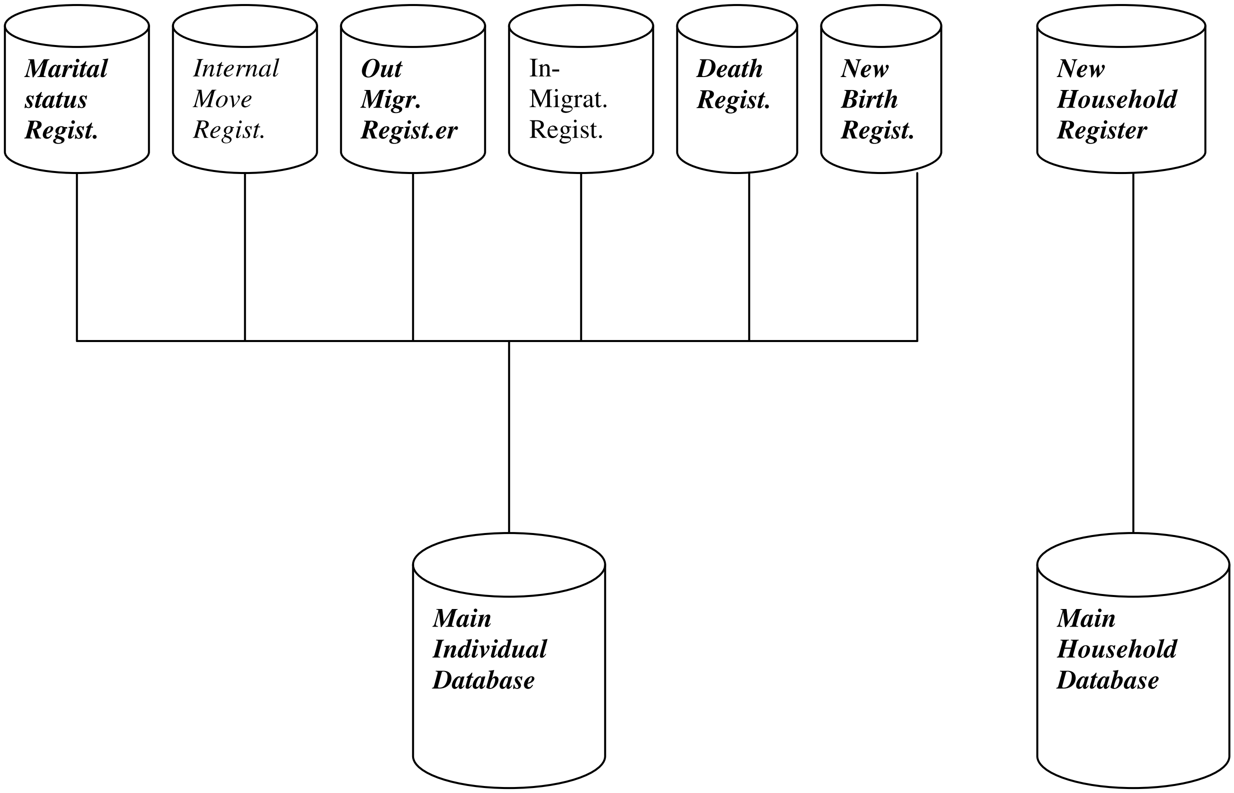


Fig 3. Conceptual design of the BRHP software system.

The software uses a menu-based approach, which (since it is running under Dos) is not mouse-friendly.

Access to the system is controlled via 7user names and passwords, operating at different levels according to the type of level required, as follows:

USERNAME	ACCESS PROVIDED[**]
READ 1	
READ 2	lists and reports - no entry or updating.
ENTER 1	
ENTER 2	as above, plus entering data to transactions.
SUPVI 1	
SUPVI 2	as above, plus updating functions.
RSCHR 1	
RSCHR 2	as above, plus record deletion.
SYSTEM	as above, plus troubleshooting mode.

Each user name is associated with its own password, which can be changed by its user. If an attempt to log in is unsuccessful (probably because of a wrong or mis-type username or password), the system closes down.

Butajira Rural Health Project

Database program Version 3.0

1. Backup menu
2. Enter event data
3. Get lists
4. Change passwords
5. Backups and file transfer
6. Update database

[**] ***A Manual of the BRHP database: Unpublished material***

```
7.  Analysis / Reports
8.  Error correction
0.  Exit to Dos
        Enter your choice:
```

The vital events menu is used for entering events collected by the enumerators in to the transaction registers, ready for latter updating to the main files. Each menu option deals with a particular type of event. In each case, checks are made in the database to ensure that an ID or household number is not assigned to a person or household that already has the same number, thus avoiding the possibility of duplicate numbers in the main database.

The updating function is very important, since it enables the main data to be updated from the transactions that have been entered. When the update process in run, a log file is automatically created in the \BUTA directory. This is a text file that contains the details of transactions that, for some reason or other, could not be processed in the updating procedure, and which therefore remain as un-posted transactions.

The Lists menu enables data to be listed and summarized from the system in a number of different ways. The options reflect functions that are likely to be useful for the day-to-day management of the data or for fieldwork, rather than for long-term analysis.

The Password Change option from the main menu does not lead to a sub-menu, but directly to a dialogue in which a user can change their password to something more secure and personalized than the initial value given. This can be done as often as required.

3.2.6 Methodological Experiences[1]

When the Butajira project started in the late 1980s, personal computers wee still in early stages of development and there was very little experience of their use in Sub-Saharan Africa, particularly in remote locations (Brass, 1987).

Yemane et. al. (1999) Wrote that initially it was unrealistic to think of computerization of the data on-site, and even at Addis Ababa University facilities for handling the data were limited at this stage. So, records were originally entered as plain text strings using a simple text editor. Whilst this served a useful function for later analysis, it did not facilitate easy interaction with the data, nor permitted real-time error checking and quality assurance procedures.

As well as the basic requirements for processing population based longitudinal surveillance data of the BRHP, a special issue arose in the Ethiopian context because of the unique calendar system in use, in which the year starts in mid-September,, has 13 months (120+30 days and one of 5 or 6) and run 2,809 days behind the international (Gregorian) system. Since field workers were much more conversant with the local Ethiopian system, this as used for dates in the field, which ever then converted as part of the data entry process. In the initial phases of the project, when data entry was in text strings, this conversion process added further difficulties (Yemane et.al., 1999)

Following a decision in 1993 that a re-census of the entire study population was called for in order to validate and update the study base, new forms for the census itself and continuing vital event surveillance thereafter were developed. Prior to the 1995 re-census, and in the light of developments in computer hardware and software in the meantime, a new software system for data entry

[1] *Yemane Berhane, et al., 1999. Establishing an epidemiological field laboratory in rural areas-potentials for public health research and interventions. The Ethiopian Journal of Health Development, vol.134, special issue.*

and storage was developed using a relational database approach. The primary aim of the new software was to facilitate the protection and safeguarding of the integrity of the data. For example, within the data checking part of this system, it was possible to include a routine for automatic conversion of data variables from the Ethiopian format on entry. It also transpired that in previous years a considerable backlog of data efforts had accumulated in the absence of checking facilities, particularly in respect of the data variables. For example, it emerged that a proportion of data of birth had been entered unconverted, resulting in some individuals being recorded as some seven years too old, and giving a very distorted age/sex distribution (Yemane et.al., 1999).

Simultaneously implementing the new software, undertaking the re-census and trying to trace accumulated errors led to a difficult period in the life of the project which took several years to work through (Yemane et.al., 1999)

Following the extensive data cleaning processes between 1995 and 1998, it proved possible, thanks to further technological advances, to commit the definitive version of the database for the first 10 years' surveillance (1987 to 1996) to a CD-ROM. This relatively simple step has proved to be very helpful to the international group of collaborators working on the analysis of this large set of data (Yemane et.al., 1999). For this research project, the datasets required for analysis by using data mining techniques, were extracted from the above-mentioned cleaned ten years' surveillance data.

CHAPTER FOUR

DATA PREPARATION AND MODEL BUILDING

This chapter details the different data mining steps that were carried out in this research work. The major steps undertaken were:

- Identifying the goal of the data mining task
- Identifying sources of pre-classified data
- Data cleaning and preprocessing
- Building and test the model

The first step in any data mining task (i.e. clear definition of the problem) has been already addressed in the first chapter of this study under section 1.2. So, it isnot discussed in this chapter. Before I start discussion on the remaining three steps as applied in this research, and for the purpose of rendering discussion in respect of each, an attempt is first made to describe the technology/tools used together with the steps involved in building a model using the tools.

In this research work, the data mining task was undertaken by using Artificial Neural Network and decision tree techniques. The specific neural network software that was used for model building & testing purposes was BrainMaker software. The researcher has also used see5 decision tree techniques to build a predictive model Thus, before going to the details of the previously specified steps that were carried out in this study, the researcher would like to give an overview of BrainMaker and See5 software.

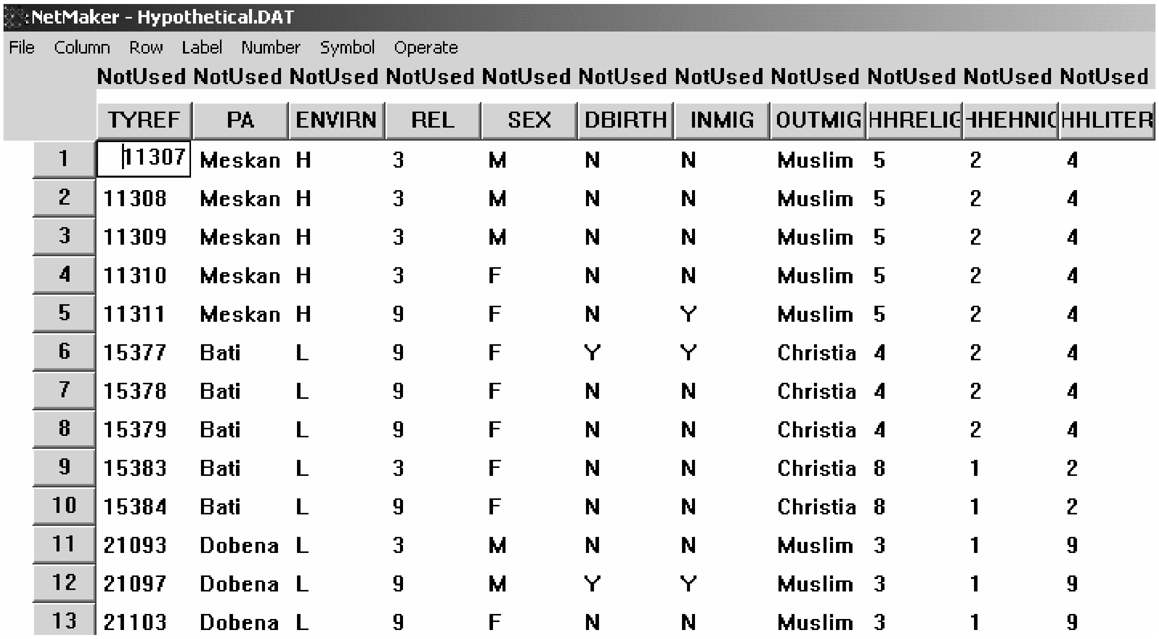
4.1 BrainMaker Neural Network Software

BrainMaker Neural Network software is developed by California Scientific Software (http://www.calsci.com/). The software is being used widely for many applications including the health care sector.

BrainMaker uses back propagation algorithm in developing a model i.e. the network is trained by presenting a set of facts over and over again. BrainMaker goes through all the training list (records) addressing each fact in turn and making necessary corrections. When the entire list of facts has been presented, BrainMaker starts over at the beginning of the list. The training process is repeated until the network gets all the facts correct or until training is interrupted. Each time through the entire fact is called a 'run' (California Scientific Software, 1998).

BrainMaker has two programs called NetMaker and BrainMaker. NetMaker makes building and training neural networks easy by importing data and automatically creating BrainMaker's neural network files. In broad terms 'NetMaker is used to create BrainMaker files, convert and perform operations on data, and graph and analyze data.' (California Scientific Software, 1998)

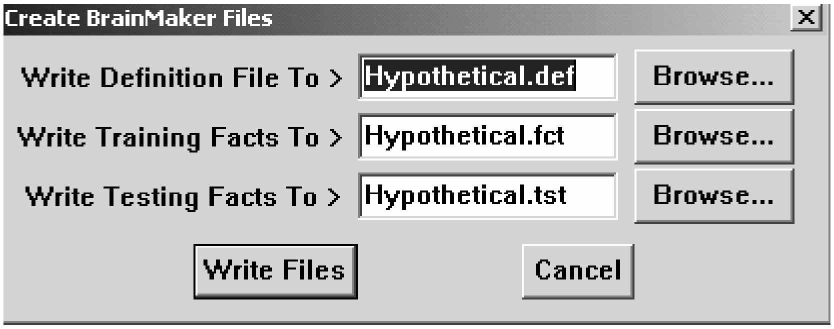
NetMaker can import data from Lotus, Excel, dBase, MetaStock, ASCII and binary files. Both numeric and text data can be accepted by NetMaker and transformed into a representation that the neural network can understand i.e. in the range of 0 to 1. The imported files are seen on NetMaker as a spreadsheet. The model developer can perform calculations and can visually analyze data while in NetMaker program. Upon completion of manipulation of data, the network builder would identify which fields would be used as inputs (independent variables) and which as pattern (dependent variable) (California Scientific Software, 1998). For this research, example of an independent variable (input) could be "Household Health" value and a pattern is the classification field that indicates whether the child is died or not. In addition, a field can be labeled as annotation, which is a label used to represent fields that are not used as independent (Input) or dependent (pattern) variables but as an identification of a particular record (California Scientific Software). For instance a ten years' reference number can be used as Annotate.



	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed	NotUsed
	TYREF	PA	ENVIRN	REL	SEX	DBIRTH	INMIG	OUTMIG	HHRELIG	HHEHNIC	HHLITER
1	11307	Meskan	H	3	M	N	N	Muslim	5	2	4
2	11308	Meskan	H	3	M	N	N	Muslim	5	2	4
3	11309	Meskan	H	3	M	N	N	Muslim	5	2	4
4	11310	Meskan	H	3	F	N	N	Muslim	5	2	4
5	11311	Meskan	H	9	F	N	Y	Muslim	5	2	4
6	15377	Bati	L	9	F	Y	Y	Christia	4	2	4
7	15378	Bati	L	9	F	N	N	Christia	4	2	4
8	15379	Bati	L	9	F	N	N	Christia	4	2	4
9	15383	Bati	L	3	F	N	N	Christia	8	1	2
10	15384	Bati	L	9	F	N	N	Christia	8	1	2
11	21093	Dobena	L	3	M	N	N	Muslim	3	1	9
12	21097	Dobena	L	9	M	Y	Y	Muslim	3	1	9
13	21103	Dobena	L	9	F	N	N	Muslim	3	1	9

Figure 4: NetMaker screen with Hypothetical Records

The prepared file is then saved with a *.dat* extension (California Scientific Software, 1998). The saved file is the basis for creating the BrainMaker files that are used to train and test a network. There are three BrainMaker files that need to be created. The first one is the definition file (*.DEF) file, which has the definition information for training such as what columns are inputs (independent variables), patterns (dependent variable) and how the information is displayed. The second is the fact file (*.FCT), which by default, constitutes 90% of the prepared data for training. The last one is the test file (*.TST), which by default has 10% of the prepared data for testing (California Scientific Software).



Create BrainMaker Files

Write Definition File To > Hypothetical.def

Write Training Facts To > Hypothetical.fct

Write Testing Facts To > Hypothetical.tst

Figure 5: Screen for the Three BrainMaker files

After the above three files are created the model developer would be ready to move to the BrainMaker program. BrainMaker program has default values for the parameters that are essential in neural network training and testing. The three most important parameters are training tolerance, learning rate and smoothing factor (momentum) (Bigus, 1996). The meanings and importance of each is described below.

Training Tolerance: Training tolerance specifies how accurate the neural network output must be to be considered correct. For example one can have an output value ranging from -20 to 30. The range for these values is 50. If a training tolerance of 0.1 is specified it would be equivalent to +-5 (10% of 50). Thus, if the training pattern is 10 and the neural network output is 15, it will be considered correct and no internal changes will be made to the network since a difference of 5 (15 minus 10) is within the training tolerance range of +-5. The software providers recommend beginning with a loose tolerance and lowering it as the network gets most of the facts correct (Bigus, 1996; California Scientific Software, 1998, Lawrence, 1994). The training tolerance concept is also applicable for testing tolerance, which is used to determine when to accept the output for the test facts as correct.

Learning Rate: In neural networks training is carried out by comparing the neural network output and the actual value and by adjusting the weights depending on the error value. Learning rate determines how big a change must be made towards the correct value i.e. do we take a giant step towards the correct value (large learning rate) or small step (small learning rate). A very high learning rate is not preferred since there would be giant oscillation as the network makes large adjustments for one pattern and another large change for the next pattern. It is recommended to lower learning rate at the beginning (Bigus, 1996).

Smoothing Factor (Momentum): This is a parameter that goes hand in hand with learning rate. The momentum parameter causes the errors from previous training patterns to be averaged together over time and added to the current error. So if the error on a single pattern forces a large change in the direction of the neural network weights, this effect can be mitigated by averaging the errors from the previous training patterns' (Bigus, 1996). In BrainMaker the default value for smoothing factor is 0.9 and the software providers state that 'adjusting the smoothing factor has not been found to reduce training time or improve prediction in every case' (California Scientific Software, 1998). Therefore, in this research undertaking, the smoothing factor value had not been changed for any of the models.

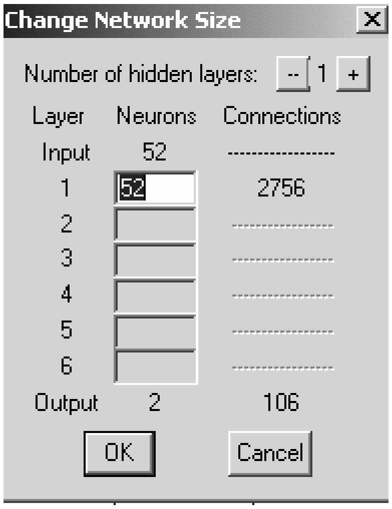
The default values, in BrainMaker, for the above parameters are:

Training tolerance	***0.1***
Learning rate	***1***
Smoothing factor	***0.9***

Table 1: Default Parameters in Brain Maker for Training Tolerance, Learning Rate and Smoothing Factor.

There are also other parameters that can be changed such as when to stop training, when to stop and make a test and how many hidden neurons and hidden layers to use etc. (California Scientific Software). The default value for number of hidden layers is one and default value for number of hidden neurons is the average of the output and input neurons but in cases where number of outputs are few the number of hidden neurons are made equal as the number of input neurons. Both these values i.e. number of hidden layers and number of hidden neurons can be changed. The software providers do not recommend change in number of hidden layers. But in the case of hidden neurons, they state that there is no hard and fast rule but recommend the following as a practical guideline (California Scientific Software, 1998).

Hidden neurons = (input neurons + Output neurons)/2



Number of hidden layers: 1

Layer	Neurons	Connections
Input	52	----------------
1	52	2756
2		----------------
3		----------------
4		----------------
5		----------------
6		----------------
Output	2	106

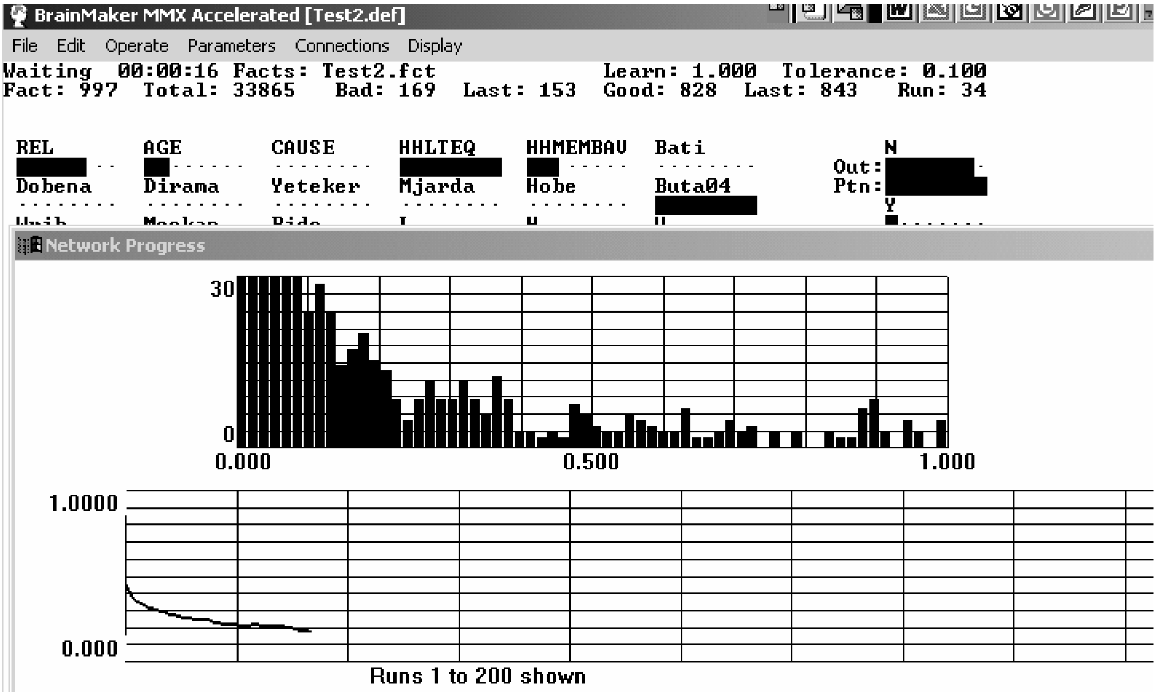
Figure 6: Screen for Network size in BrainMaker

The above shows a description of a network size in BrainMaker. This network size was created for the hypothetical data put under figure 5 above.

The other concept in the above network size is the connection numbers. The screen shows that there are 2756 connections between input and hidden neurons. Each input neuron is connected to all the hidden neurons, which bring the connection numbers to 2704 (52*52). But as seen from above the total connections between the input and hidden neurons are 2756. This is due to the presence of a bias unit. In a back propagation training a bias unit is added to the input and hidden layers. The values of the bias units is always one and they are important in avoiding the possibility of having a network where it would be impossible to change the weights in cases where all inputs are zero. In the above network, the presence of the bias unit in the input layer increases the number of connections to 2756. Also the connections between hidden and output neurons are 104 (2*52) without considering the bias unit. But since there is one bias unit in the hidden layer, the total connection between the hidden and output nodes becomes 106. In summary the above network has 2756 weights connecting the input to the hidden nodes and there are 106 weights connecting the hidden to the output nodes.

The other concept that is considered important in neural networks is which function to use. BrainMaker supports four kinds of functions namely sigmoid, threshold, step and linear functions. However, most popular function is the sigmoid function. And the software providers also state that they "... have seen no problems which train fundamentally better with anything other than the standard sigmoid transfer function" (California Scientific Software, 1998). Therefore, for all the trainings in this research work it is the sigmoid function that was used.

To start training the command is 'Operate/Train Network' being in BrainMaker window. While training progresses statistical information are provided on the screen such as which fact the BrainMaker is processing at a specific time, the number of facts which met and did not meet the training tolerance, the number of run (epoch) etc. There are also two graphs that display the progress of the training. The first is a histogram that shows the distribution of error over an entire run. The horizontal axis represents the error level and the vertical axis signifies the number of output values at the level. As training progresses and fewer facts are classified as incorrect, the bars (solid boxes) move to the left. The second graph shows the progress of the error rate as network trains. In this graph the horizontal axis shows the run number while the vertical axis represent the overall error level (RMS error). For a good training the error value (RMS error) would decrease as the number of runs increases (California Scientific Software, 1998).



```
Waiting  00:00:16 Facts: Test2.fct            Learn: 1.000  Tolerance: 0.100
Fact: 997  Total: 33865   Bad: 169  Last: 153  Good: 828  Last: 843   Run: 34

 REL        AGE        CAUSE      HHLTEQ     HHMEMBAV   Bati            N
                                                                  Out:
 Dobena     Dirama     Yeteker    Mjarda     Hobe       Buta04    Ptn:
                                                                      Y
```

Figure 7: BrainMaker screen while in Training

Training can be stopped at any time before the instruction to stop training is met. And the model developer can test the network and save it if it results in an acceptable prediction rate. Or it is possible to wait until training stops and test the model (network). The default for stopping training is when the incorrect classification of facts in a run (epoch) becomes zero. The software vendors discuss that better network (model) is usually obtained before the criteria for stopping training are met (Lawrence, 1994). Therefore, it is advised that the model developer periodically save a network.

BrainMaker Neural Network Software has the facility whereby data are classified into two sets i.e. training and testing set. The software automatically puts aside 10% of the records into a testing file. These data would be used to test the accuracy of the network and would not be seen by the network during training (California Scientific Software, 1998). This step i.e. dividing of the data into training and testing records is carried out for every network that is experimented with. Dividing of data set into training and testing sets is also the approach suggested for model building by different writers such as Bigus (1996) and Lawrence (1994).

Among the saved network the one with the best result would be selected and a running fact file would be developed that can be used in predicting future records (California Scientific Software).

4.2 See5 Decision Tree Software

Since See5 decision tree software was the other data mining technique/tool that has been used to build models with better performance particularly in determining the most important factor(s) for child mortality at the district of Butajira, a brief introduction about this software is presented as follows.

See5 is a windows based data mining software that analyses data to produce decision trees and/or Rulesets that relate a case's class to the values of its attributes. See5 uses different application files to define classes and attributes, to describe the cases to be analyzed, to provide new cases to test the classifiers produced by See5, and specify misclassification costs or penalties (See5: An informal Tutorial http://www.rulequest.com/see5-win.html).

Every file associated with an application has a name of the form *filestem.extension* where filestem identifies the application and extension defines the file's type or function. Two files are essential for all See5 applications and there are three further optional files, each identified by its extension. The first essential file is the names file (e.g. in our case *mortality.names*) that describes the attributes and classes.

In this research project, the names file, which describes the attributes selected to build the models and their classes, was created using word pad (*See Annex2*).

The second essential file for any See5 application is the applications data file (e.g. *mortality.data*). This file contains the cases that will be analyzed in order to produce the classifier. Each case is represented by a separate entry in the file. The entry for each case consists of one or more lines that give the values for all explicitly defined attributes. Values are separated by commas and optionally terminated by a period. For the purpose of this research project, the data file (*testdata.data*) prepared for see5 consisted of all the cases that were used to build the neural network models. For example, the first three cases from the data file prepared for this research work (*testdat.data*) looks as follows:

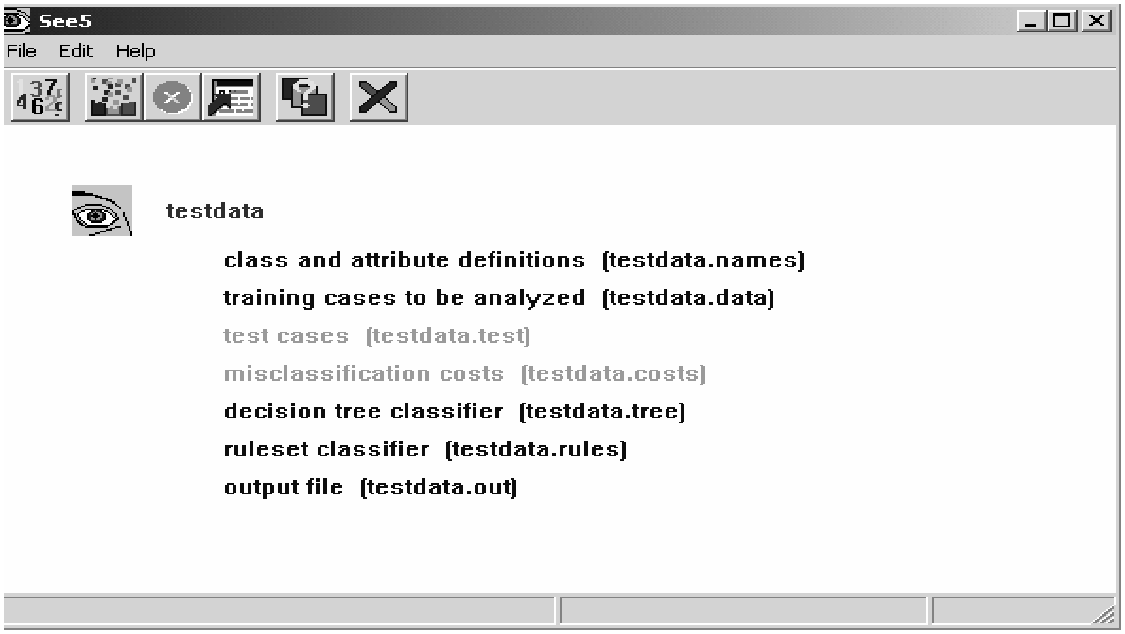
```
Meskan,H,4,F,N,N,Muslim,Meskan,Literate,7,TradHeal,Protwell,Tukul,Y,N,
Died,10093.
Meskan,H,1,F,N,N,Muslim,Meskan,Illitrate,9,TradHeal,River,Tukul,Y,N,
Died,10103.
Meskan,H,2,M,N,N,Muslim,Meskan,Illitrate,5.5,GovtHCU,River,Tukul,Y,N,
Died,10115.
```

After the cases in the data file have been analyzed, the predictive accuracy of the classifier generated from them can be estimated by evaluating it on new cases.
To do this, *See5* uses a third kind of file, which consists of new test cases on which the classifier can be evaluated in terms of the error rate on new cases.. This file is optional and, if used, has exactly the same format as the data file (See5: An informal tutorial, http://www.rulequest.com/see5-win.html). In this research work, a separate test file was not prepared. Rather, the performance of the classifier was evaluated by using the “ Use Sample” option provided by See5. Thus, by invoking this option, samples cases were randomly selected from the data file.

Another optional file used by See5 is the cases file. This file differs from a test file only in allowing the cases’ classes to be unknown (?). The cases file is used primarily with the cross-referencing procedure and public source code, and as such was not used for this research work.

The last type of file, the costs file, is also optional file and sets out differential misclassification costs. By using this file See5 allows different misclassification costs to be associated with each combination of real class and predicted class.

Once the names, data, and optional files have been set up, everything is ready to use See5. To employ See5 for classification, testing, prediction, and cross-referencing tasks, there is a user interface (shown below), which consists of buttons and menu commands. This main window has six buttons on its toolbar where each button is used to perform different tasks.

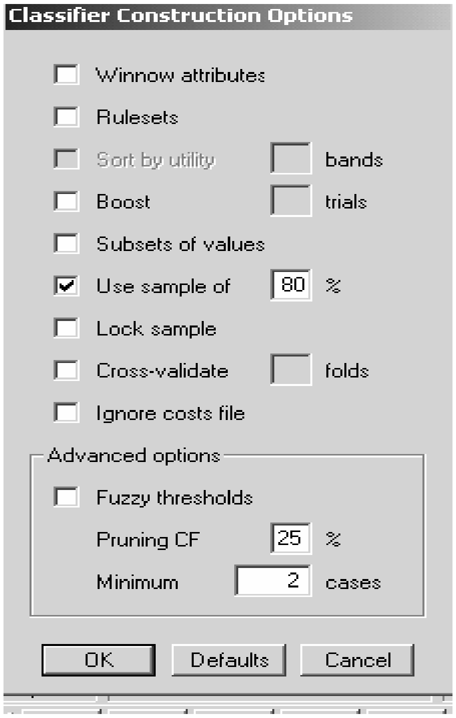


testdata

class and attribute definitions [testdata.names]
training cases to be analyzed [testdata.data]
test cases [testdata.test]
misclassification costs [testdata.costs]
decision tree classifier [testdata.tree]
ruleset classifier [testdata.rules]
output file [testdata.out]

Figure8: See5 Main Window

The first step is then to locate the data using the locate data button on the toolbar (or the corresponding selection from the File menu). After the application data file is located, the classifiers will be constructed. If Construct classifier button of See5's main window is used, it provides several options that affect the type of classifier that See5 produces and the way that it is constructed. The construct classifier button on the toolbar displays a dialog box, which consists of construction options displayed on the following diagram.



Classifier Construction Options

□ Winnow attributes

□ Rulesets

□ Sort by utility [ ] bands

□ Boost [ ] trials

□ Subsets of values

☑ Use sample of [80] %

□ Lock sample

□ Cross-validate [ ] folds

□ Ignore costs file

Advanced options

□ Fuzzy thresholds

Pruning CF [25] %

Minimum [2] cases

Figure 9: Classifier construction Options dialog box

When See5 is invoked with the default values of all options, it constructs a decision tree output. The evaluation of the decision tree on training and test cases is also provided with the output. In this research work, See5 was employed to build classifiers by using the default values as well as options such as ruleset and adaptive boosting. After the models were built by using the various options provided by See5 software, the prediction performance of some of those models was also evaluated by using new cases (validation data set) that were not used to build the model.

The above in brief explains the steps involved in building a model using both BrainMaker Neural Network and See5 decision tree software. Now we turn to the steps that were followed in this research work.

4.3 Identifying Sources of Pre-classified Data

As it has been stated under section 1.2 of the first chapter, the primary goal of this research work was to assess how well a predictive model built by using data mining techniques would perform in predicting the risk of mortality among children based solely upon epidemiological datasets consisting of demographic, parental, environmental, public health and epidemiological history factors, without using diagnostic tests or physical exam findings. This type of prediction model might have an application outside of clinical settings to support health care workers in taking preventive actions to reduce child mortality in the region, as well as to assist health care planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at improving child survival in the region. Thus, the results of this data mining process can be used to distinguish children at high risk of mortality from those with low risk of mortality.

Thus, for any data mining task, the primary requirement is availability of data in any format (Berry and Linoff, 1997). And in most cases the ideal source of such data is the corporate data warehouse (where data warehouse refers to the collection of data from many different sources and its storage in a common format with consistent definitions for keys and fields).

Especially, in supervised learning systems, we use pre-classified historical data (past data) to build a model of the future. For example, in relation to this research work, in order to distinguish children who are likely to have high risk of morality from those with low risk of morality, the model should be trained using historical data collected from each class of children (Died and Alive) which is then used to build a predictive model that can predict the survival probability of new cases. Implicit in this model is the idea that we can tell what has happened in the past to predict the future.

In this research, in order to identify sources of historical data collected about children by BRHP, discussions were held with the system administrator of the project's main database and other academic staff members of the Community Health Department of AAU. As a result of these discussions, the researcher identified two main sources of data (both in electronic format) that can be used for this research project.

The first source of data identified was the main database of the BRHP project. This database stores a comprehensive population-based data gathered about all individuals registered in all the ten villages of the study area for the last 15 years. This database contains complete demographic and health related data gathered through census and epidemiological surveillance system. The surveillance system provides a means for on-going registration of vital events (such as birth, death, cause of death, marriage, new-household, out-migration, in-migration etc.). Such data is collected monthly by visiting each household and by checking and recording new events occurred. Through the surveillance system, the status of each individual in every household is checked and recorded on a demographic surveillance sheet, and then this data will be transformed into transaction files, and finally permanently stored in the main database. Information like basic demographic characteristics of each individual, housing condition, and health care utilization behavior of each household are recorded at entry into the surveillance system and during any re-census process.

Data entry into the main database is performed at the Department of Community Health, AAU using a specifically designed dBase IV relational database software. So, during data entry, the data in each event form is entered into a separate transaction file by data entry clerks and then from the transaction files it will be transferred into the main database for permanent storage.

This database is a database which stores every new data copied from transaction files, and it is updated every month to include new events (such as death, birth, migration, new household, etc.) occurred in the study population of the Butajira project.

The second source of data identified as a potential source of data for this research work was the separate ten years' surveillance data set (1987 - 1996) which was created as a result of the data cleaning activities carried out on the main database of the project between 1995 and 1998. This data set is a definitive version of the main database, but it is a separate data set, which was purposefully cleaned and stored to assist international group of collaborators working on the analysis of this large set of data by providing a separate data set to wok with. This data set contains a total of 64,077 records gathered through out the ten years' period.

Since this ten years' data set is a data set that has been stored separately from the main database to assist the researchers by providing a separate data set to work with, it was this separately stored surveillance data set which was selected and used as a source of data needed for this research project.

The ten years' surveillance data set was originally stored in EPI6 format (which is a statistical package usually used for the analysis of medical data). This data set was organized in rows and columns where each column represents an attribute (variable) and each row stands for a single record of an individual. In this data set a total of 77 attributes (columns) and 64,077 records (rows) were identified.

The researcher took a zipped copy of this original data set and all the necessary operations of data selection and data preparation were carried out on this data set. Selecting the target data set required for this research work was conducted by using the selection facility provided by EPI6 software. Thus, the "SELECT" command was used to select records about children, whose date of birth is greater than or equal to January 1, 1987. The selected data set consisted of, a total of 15,667 records of children who were born between January 1, 1987 to

December 31, 1996 in all the ten villages of the BRHP study area. Of these records of live births, 2,330 records were about children who were born alive but died after some time. The following table shows the distribution of those died children throughout the ten-year period in all the ten villages of the BRHP study area.

Table2: Distribution of died children in the ten villages of the Butajira project

No.	**PA(Peasant Assn.)**	**Environment**	**# of Children Died**
1	Meskan	H	185
2	Bido	H	301
3	Dirama	H	431
4	Wrib	H	150
5	Yeteker	H	160
6	Bati	L	203
7	Dobena	L	209
8	Mjarda	L	162
9	Hobe	L	261
10	Buta04	U	187
		Total	2330

Following the successful selection of the required data set from the original ten years' data, the next important issue considered by the researcher was importing the selected data set, which was in EPI6 document format into Ms-Excel format.

Although the original data set was organized in rows and columns using EPI6 statistical software, it was not possible to import this data set as it is into the data mining software that were used in this research project. Since BrainMaker software has a facility to import data from Lotus, Excel, dBASE, Metastock, AscII and binary files, the researcher decided to import the original data set into MS-

Excel format, which will then be imported to BrainMaker software. As a result of this attempt, the researcher has successfully imported the selected data set into Excel format.

In order to build a model that can be used to predict child mortality risks in the study area based upon environmental, parental, and health related factors, a data set consisting of both classes' of children (Died, and Alive) was created and prepared for analysis. Since this research work is a preliminary work intended to assess the applicability of data mining technology based on epidemiological data sets, models were built by samples taken from the original ten years surveillance data Another reason for not considering the entire dataset was shortage of time particularly to perform data cleaning and pre-processing tasks. Thus, for the purpose of this research work, a sample data set consisting of records from each class of children was selected and used to build both the neural network and decision tree models. To select samples of died children from each PA (Peasant Association), a stratified random sampling method was used. As it is presented under table 2 above, the study area is divided into ten different villages (PAs). To select proportional samples of died children from each PA, the ten villages were stratified into 10 strata and from each strata (PA), 25% sample records of died children were selected randomly. This process has resulted a total of 560 died cases to be selected from this class. In order to insure an equal representation of events (death and survival), 540 sample records about Alive children were also selected randomly from each of the ten-villages. Selection of sample records about survival cases was conducted by taking into consideration the number of died cases taken from each village (PA).

So, to build and test the neural network model required for this research work, a sample data set which contains a total of 1100 sample records about both Died & Alive children was created and prepared for analysis.

The number of sample records selected about both Died and Alive cases from the ten villages of the BRHP study area is summarized in the following table.

Table 3: Number of sample records selected from each PA

PA (Peasant Association)	**DIED**	**ALIVE**	**Number of Records Selected**
Meskan	47	45	92
Bido	76	71	147
Dirama	102	98	200
Wrib	38	38	76
Yeteker	40	40	80
Bati	48	47	95
Dobena	52	50	102
Mjarda	42	40	82
Hobe	66	64	130
Buta04	49	47	96
Total	560	540	1100

Determining the total number of examples needed to train neural networks is an application dependent task. There is no precise method that can be used as a standard. However, the most important issue to be considered is that, the more facts and the fewer hidden neurons we use, the better will be the performance of the network. Moving too few facts or too many hiddens can cause the network to "memorize". When this happens, the network performs well during training but tests poorly.

Although there is no hard and fast rule that can be applied to determine the total number of training facts, BrainMaker software vendors have suggested certain guidelines that can be used to determine the total number of training facts and number of hidden neurons required to build a good neural network.

According to the guideline suggested by the venders of BrainMaker software, the number of training facts needed to build the required neural network model should lie between two and ten times the number of input, hidden, and output neurons in the network.

Training facts = 2 * (inputs+ hiddens+ outputs) to
10 * (inputs + hiddens + outputs)

Thus, the number of training facts and hidden neurons required for this research work was determined based on the suggestions proposed by the vendors of BrainMaker software. For example, in this research work, if the first training trial is considered, there are 50 input, 50 hidden, and 2 output neurons. So, based on the above guideline, the number of training facts required for can be determined as follows:

Training Facts = 2*(50+50+2) to 10*(50+50+2)
= 204 to 1020

Thus, for this trial the total number of facts used for training should lie between 204 and 1020. For that particular trial and other models built in this research work, a total of 990 training cases were used to build models which was actually closer to the highest value identified by using the suggestion made by the vendors of BrainMaker software.

Feature Selection:

As it is described previously, the original data set from which the target data set for this research work has been selected consisted of a total of 77 attributes. However, based on expert's opinion, of these attributes, only some of them were considered as relevant (influential) for the specific learning task to be undertaken in this research work.

Discussing the importance of selecting relevant features (attributes) in any data mining task, Liu and Motoda (1998) wrote that “the abundance of potential features constitute a serious obstacle to the efficiency of most learning algorithms. Popular methods such as k-nearest neighbor, C4.5, and back propagation are slowed down by the presence of many features, especially if most of these features are redundant and irrelevant to the learning task.” The authors further stated that some algorithms may be confused by irrelevant or nosily attributes and construct poor classifiers. Therefore, eliminating some attributes, which are assumed to be irrelevant to build the model can increase the accuracy of the classifier, save the computational time, and simplify results obtained.

It was by considering the importance of selecting relevant features (attributes) in increasing the efficiency of the algorithm that the researcher opted to apply dimensionality reduction on the data set created for analysis by selecting the minimum set of features (attributes) that are associated with the learning task.

As a result of the dimensionality reduction process, a total of 18 candidate attributes were selected as relevant features to build and test the required neural network model. Selection of these relevant features was conducted in consultation with public health specialists, pediatricians, and researchers who have good knowledge and experience on the data set of BRHP database.

The following table gives a brief description of the candidate attributes selected as relevant features to build and test the required models.

Table 4: Attributes selected from the original data file

No.	Attribute Name	Description	Values
1	TYREF	A unique reference number given to each child	A five digit serial number
2	PA (Peasant Association)	The name of the village in which the child is registered	10 unique categorical values.
3	ENVIR	The climate of the child's village	3 unique categorical values
4	SEX	The gender of the child	2 symbolic values
5	DBIRTH	The month/data/ year in which the child was born	Date of the form MM/DD/YY
6	HHREIG	Religion of the child's parents	3 unique categorical values
7	HHETHNIC	The ethnic origin of the child's parents	8 unique categorical values
8	HHLITER	The level of education of the child's parents	3 unique categorical values
9	NHHMEMBAVE	The average number of family members in the child's family	Continuous numerical value.
10	HHHEALTH	Main source of water supply for the child's family	7 unique categorical values
11	HHWATER	Main source of water supply for the child's family	7 unique categorical values
12	HHROOF	The roof of the house in which the child's family lives	2 unique categorical values (i.e. Tukul or Corrugated Iron
13	HHLIVEST	Availability of cattle inside the house of the child's family	Yes/No values
14	WINDOWS	Availability of windows in the House of the child's family	Yes/No Values
15	Relation	Relation of the child with the head of the household	5 unique categorical values
16	Age	The age of the child	Continuous
17	INMIG	In migration of the child's family	Yes/No values
18	OUTMIG	Out-migration of the child's family	Yes/No values

The above selected fields were selected with different purposes in mind. Some of those features were selected to be used as an independent (input), while others were selected to determine other important variables (such as date of birth to compute age).

The dependent (output) variable i.e. status of the child was used to classify children as Died and Alive. Based on the outcome of the dependent variable, the contribution (influence) of predictor (input) variables for child mortality will be evaluated.

Fields that were available in the original ten years' surveillance data set but that were considered as irreverent for the learning task of this research project include: "Number of Household", "Date of In migration", "Date of Entry ", "Date out Migration", "Cause of Death" etc.

Since the data source selected for this research project consisted of the separate ten years surveillance data, records collected after January 1997 were not considered in this research project. As it is already discussed previously, past records of children who were born between January 1987 and December 1996 were used to build a model, which can predict the likelihood that a given child would die or survive.

4.4 Preparing Data for Analysis

As it is described in the data mining methodology of Berry and Linoff (1997), data cleansing and data preparation is the second step of any data mining task. In particular, data collected from different sources has to be massaged into a form that will allow the data mining tools to be used to best advantage. This process of data cleaning and preprocessing is highly dependent on the technique (method) to be employed. In this research project, artificial neural networks and decision tree data mining techniques were used to build a predictive model. So, data preparation was conducted by taking into consideration the requirements of both the neural network (BrainMaker) and decision tree (See5) techniques.

Therefore, in preparing the data set into a form that is convenient for BrainMaker and See5 software, the following data cleaning and preprocessing steps were undertaken on the data set created for analysis. Models Built by using BrainMaker are presented under section 4.5, and for the See5 decision tree models, section 4.6 will give the details.

4.4.1 Data Transformation and Reformatting

In general, how you represent your data is often crucial to the success or failure of your neural network project. There are some common areas of confusion that should be avoided during the data preparation stage of the data mining task. Particularly, when some data appear as numbers, they may deceive us. So, if numbers represent unique concepts and not values within a continuous range of some quantity or rating, they should be converted into symbols or separate columns (California Scientific Software, 1998).

In the original data set created for this data mining task, the values of some attributes were represented using numerical codes. Such numerical codes that are used to represent unique concepts had to be converted into symbols or columns to avoid any confusion during training and testing of the model as follows:

Decoding the "PA" code: in the original data file, the attribute "PA" which stands for "Peasant Association" was represented using ten different alphanumerical codes. So, for the purpose of this research work, those numerical codes were converted into their respective symbolic values (the name of the PA). The following table shows the PA code used in the original data file and the reformatted name of the PA.

Table 5: Actual Values of each PA code and the original numeric code.

Original "PA" Code	Reformatted Value Used
005	Meskan
007	Bati
008	Dobena
011	Bido
04B	Dirama
06A	Yeteker
06B	Wrib
09A	Mjarda
09B	Hobe
K04	Buta04

Decoding the "HHRELIG" code: in the original data set the variable "HHRELIG" stands for "House hold Religion" and its values were represented using three different numerical codes one for each value. For this attribute, the original numeric codes were converted into the respective symbolic values (the religion of the household). The following table shows the numeric code used in the original data set and the actual symbolic value used to build and train the required model.

Table 6: Original "HHRELIG" Code and the reformatted actual value

Original "HHRELIG" Code	Reformatted Symbolic Value
1	Christian
2	Muslim
3	Other

Decoding the "HHETHNIC" code: in the original data set, the variable "HHETHNIC" stands for "House Hold Ethnicity" (i.e. the ethnic origin of the household) and its values wee represented using eight different numerical codes where each code represents one unique symbolic value. So, for this attribute, the numeric codes used in the original data set were converted into their respective symbolic values.

The following table shows the original numeric codes used to represent the values of "HHETHNIC" attribute, and the reformatted numeric value for each code.

Table7: Original "HHETHNIC" code and the decoded symbolic value

Original HHETHNIC Code	**Reformatted Symbolic Values**
1	Sodo
2	Dobi
3	Meskan
4	Maraku
5	Silti
6	Amhara
7	Oromo
8	Other

Decoding the "HHLITER" code: this variable stands for "Household Literacy" (i.e. the level of education of the household) and in the original data file its values were represented by using two different numeric codes where each code represents one unique categorical value. These numeric codes wee there fore converted into their actual values at this stage this research work.

The following table shows the original numeric codes used to represent the values of "HHLITER" attribute, and the decoded actual values for each code.

Table 8: Original "HHLITER" code and the decoded symbolic value

Original HHLITER Code	**Decoded Symbolic Value**
1	Literate
2	Illiterate

Decoding the "HHEALTH" code: this variable is used to store information about the source of health care services and facilities for a given household. In the original ten years' surveillance data set, the values of this variable were represented using ten (1-10) single digit numeric codes where each single digit numeric digit represents one symbolic value related to the source of health for a given household. So, for the purpose of this research work, the numeric codes in the target data set created for analysis were converted into their respective actual values.

Table 9: Original "HHHEALTH" code and the news decoded values

Original HHHEALTH Code	**Decoded Value**
1	Gov't Health care unit
2	CHA (Clinic Health Assistant)
3	Pharmacy
4	Traditional Health Assistant (THA)
5	
6	Self Treats
7	Private Clinic
8	Health Post

9	Nothing
10	Other

Decoding the "HHWATER" code: This variable is used to maintain information about the source of drinking water for each household registered by the BRHP. In the original data file, the values for this attribute were represented using seven (1-7) single digit numerical codes. Therefore, for this variable as well, these numerical codes were converted into their respective actual values. The following table shows the original codes and the respective values for those codes.

Table 10: Original "HHWATER" codes and the decoded values.

Original Code	**Decoded Values**
1	River
2	Protected Well
3	Unprotected well
4	Lake
5	Pond
6	Pipe
7	Other

Decoding "HHROOF" code: This variable is used to gather information about the type of material used to cover the roof of the house of a given household. For this variable two single digit (1-2) numeric codes were used in the original data set. The actual values required for this variable were decoded as follows:

Table11:Original codes for "HHROOF" and the decoded actual values

Original Code	**Decoded Values**
1	Thatched (Tukul)
2	Corrugated Iron

In order to create new variables from the existing ones and to reformat the original values of some attributes in the sample data set selected for analysis, the following data transformation and reformatting operations were employed.

Creating the age attribute: in the original data set age is not included as an attribute. However, using age as one input (predictor) variable can help to categorize children into different age groups. This categorization would help to identify mortality patterns among children with different age groups. So, the age attribute was created from the date of birth attribute.

4.4.2 Handing Variables with Missing Values

In order to handle the problem of variables with missing values in the data set created for a given data mining task, there are different approaches recommended for use. For this specific research work, the approach suggested by Two crows corporation (1998) was adopted to represent missing values identified in the target data set.

- For continuous variables (variables with numerical values such as age), missing values had to be replaced with the mean value for that field (Two Crows Corporation; 1999). Thus, based on this suggestion, for all numeric variables used in this research work, any missing value was replaced by the mean value of those fields. So, for the variable "Average number of house hold member ", all missing values were replaced by the mean value of this variable. The average value for this variable was 4.6.

- To handle the problem of missing values for categorical variables, the suggestion is to group such fields into nominal and ordinal variables, and to take the median for ordinal variables and the modal value for nominal variables. The ordinal variables are categorical variables whose values have meaningful order (such as: High, Medium, Low), where as nominal variables refer to the categorical variables whose values are unordered.

Therefore, in the data set created for this research work, categorical variables whose missing values were handled in accordance with the above suggestion consisted of the following: "HHRELIGION', 'HHETHENIC', 'HHLITERACY', 'HHHEALTH', 'HHWATER', 'HHROOF', and 'HHLIVEST'. Since all the above fields are nominal variables, for any missing value in those fields, the modal (most frequent) value was used. The modal values for the above nominal categorical variables are given in the following table.

Table 12: Modal Values to represent Missing Values

Field Name	***Modal Value***
HHRELIG	Muslim
HHETHNIC	Silti
HHLITERAC	Illiterate
HHHEALTH	Health post
HHWATER	River
HHLIVEST	Yes
HHROOF	Tukul

4.4.3. Preparing the Data set into a form that is Acceptable to the Neural Network.

Neural networks accept values only in the range of 0 to 1 or -1 to +1. Therefore, all values in the data set have to be represented with values ranging from 0 to 1 or -1 to 1 to be analyzed by the neural network software chosen.

BrainMaker software which was used for this research project provides a facility to automatically transform both symbolic and numeric values into a form that can be understood by the neural network i.e. in the range of 0 to 1 as its default range it also offers an opportunity to change the default range and to transform values in the range of -1 to 1. As discussed in the first section of this chapter, one input neuron is assigned for every numeric variable, and the values in the numeric field are scaled down to the range of 0 to 1 or -1 to 1 if the default ranges are changed.

In the case of categorical variables, the number of input neurons is equal to the number of unique values in that field. For instance, in the "HHHEALTH" field, there are 10 unique symbols (values), which are allowable values for this field. These are "GovtHCU", "CHA", "pharmacy", "Traditional Health Attendant", "Self Treat", "Health Post", "privateClinic", "Nothing", and "Other". Therefore, nine input neurons are assigned for this field and the presence of a symbol in the input data turns that neuron on (California Scientific Software, 1998).

4.5 Models Built Using BrainMaker Software

To build the neural network model, the first task performed at this stage of the data mining process was importing the cleaned and prepared data set, which was in Excel format, into NetMaker. (See Annex 1 for sample dataset prepared). This prepared data set consisted of 1100 sample records about infants and children extracted from the original ten years' surveillance data set of the BRHP.

As it was explained in the first section of this chapter, BrainMaker software has two independent programs: Netmaker and BrainMaker. Netmaker program provides a facility to import data files from dBase, Excel, Lotus etc., to carry out manipulations on the imported data file, and to create BrainMaker files. BrainMaker program has also facilities to train and test networks (models) and to create running fact files to build the model required.

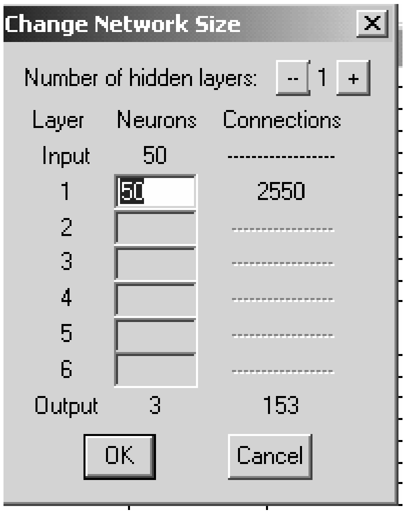
Thus, after the data set in Excel format was imported into NetMaker, all the necessary data manipulations (such as defining the *input*, *annotate*, and *output* or pattern attributes), creating BrainMaker files etc. were carried out using NetMaker's facilities. Then, BrainMaker program was used to train, test and create running fact files.

The first training trial was conducted by using all the 16 input attributes that were selected during the data preparation phase. Those variables that were used to build the first network (model) were: "PA", "SEX", "AGE", "ENVIRN", "HHRELIG", "HHETHNIC", ""HHLITERAC", "HHMEMBAVE", "REL", "INMIG", "OUTMIG", "HHHEALTH", "HHWATER", "HHROOF", "WINDOWS", and "HHLIVEST".

For the initial training trail, all the default parameters provided by BrainMaker (i.e., *Training Tolerance, Learning Rate, Smoothing Factor, Number of Hidden Neurons,* etc.) were accepted and used as they were. The default parameters used for this initial trial has the following values:

Training Tolerance	***0.1***
Learning Rate	***1.0***
Smoothing Factor	***0.9***

The default number of hidden neurons provided by BrainMaker for this trial was 50, which is set equal to the input neurons by default. This number represents the number of values used for each of the above attributes selected to build this model. As it can be seen above list of variables used to build this model, out of the i6 input attributes, 14 of them were categorical and only two of them were numeric attributes. For each value of categorical variable a single input node will be assigned, where as for each numeric variable a single input node is assigned. Thus, the number of input and hidden nodes in this network is the result of the sum of the input nodes assigned to the values of categorical and numeric variables. This initial network has the following network size.



Number of hidden layers: 1

Layer	Neurons	Connections
Input	50	-----------------
1	50	2550
2		-----------------
3		-----------------
4		-----------------
5		-----------------
6		-----------------
Output	3	153

Then, upon the command “***operate/Train Network***”, training resumed. In this particular run, after about 806 facts it became difficult for the network to learn the other 183 facts and the network progress display facilities that are provided by BrainMaker to visualize the network progress during training indicated that there was no improvement (i.e. the distribution of errors and the RMS error value stopped to decrease). This may be an indication that more hidden neurons are needed, there may be a problem with data representation, or your maximum/minimum range may need attention” (California Scientific Software, 1998). Therefore, in order to trace problems that forced the network to have a trouble in learning all the training facts, this initial training attempt had to be abandoned. However, before altering/adjusting parameters and building another network, this network had been saved and analysis of this network showed a precision rate of 77% (85 of the110 test facts were classified correctly) at testing tolerance of 0.4. To see the performance of this network with a more tight testing tolerance, the default testing tolerance was changed into 0.2, and the network showed a precision of 74% (81 of the 110 facts were classified correctly). And upon a decision to stop training, another network was saved. This network had a precision rate of 77% and 73% at testing tolerances of 0.4 and 0.2 respectively.

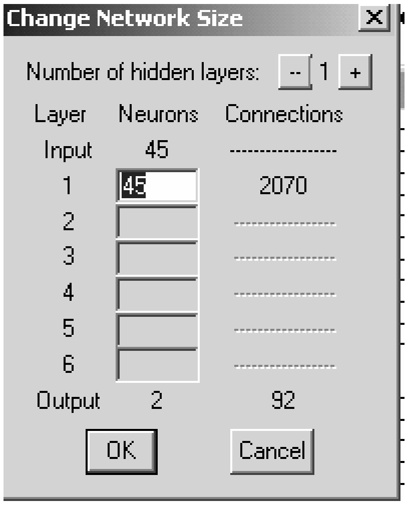
To address the problem of the above network, the first option considered by the researcher was to train and test a network by varying the default parameters. The parameters were changed in accordance with suggestions provided by various writers (such as Berry and Linoff,1997; Bigus,1996) . Thus, the default parameters that were used to build the initial model were changed as follows:

Training Tolerance	***0.3***
Learning Rate	***0.6***
Smoothing Factor	***0.9***
Number of hidden neurons	***26***

This network also failed to converge, but was still tested at different intervals and resulted in a less performing network than the above. Three of the networks tested with the accuracy of 77%, 65%, and 71% at a loose testing tolerance of 0.4. And at a tighter testing tolerance of 0.2, the networks resulted in a precision of 62%, 61%, and 64% respectively.

Following the previous attempt, an attempt was made to test by varying the composition of the input variables. This was done in consultation with experts in the area who had better knowledge and experience as to which variables are more determinant and essential to predict child mortality. However, a better result was not obtained. One of these trials was a network, which had 13 of the 16 input variables used to build the previous model. The excluded three variables were

“Relation”, “in migration”, and “Out migration”. To build this network, the default parameters were used. This network has the following network size.



Change Network Size

Number of hidden layers: 1

Layer	Neurons	Connections
Input	45	----------------
1	45	2070
2		----------------
3		----------------
4		----------------
5		----------------
6		----------------
Output	2	92

Several networks were saved during training and the highest precision obtained from these networks was 72% at testing tolerance of 0.4.

Since all the previous test results were encouraging, the researcher continued the experiment by considering various options suggested to improve the performance of neural network models. For example, the vendors of BrainMaker software suggested, “when your network got a sufficient number of training facts correct and performed fairly well in testing, try to build a network by changing the training tolerance, learning rate, shuffling the data, reducing the number of hidden neurons, and by adding a noise” (California Scientific Software, 1998).

So, the researcher checked if there are outliers for the numeric variables used. Upon visualization of the data using a histogram, “age” and “average house hold member”, variables were checked for outliers and the researcher identified no outliers in those variables. The histogram displayed for those variables is presented as follows:

Figure 10: Column histogram for the Age attribute

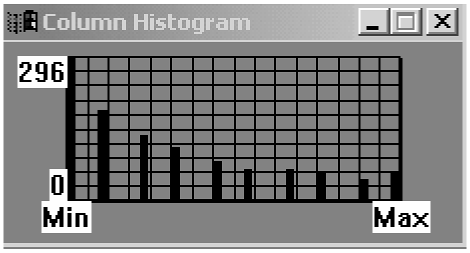
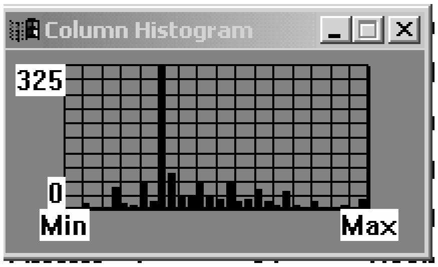


Figure 11: column Histogram for "HHMEMBAVE"



As it observed from figure 10, there is no an outlier in the age field. That means all the values are well distributed (i.e., there is no single value which dominates the others). For figure 11 as well, an outlier is not observed since the values are distributed between the maximum and the minimum.

Another option considered by the researcher was reshuffling the order of the facts so that to change the order of the facts and to avoid grouping of similar facts together. In relation to this, the vendors of BrainMaker software described that neural networks might have a trouble grasping the overall solution if the facts are grouped by similar inputs or outputs (California Scientific Software, 1998). In such cases it is recommended that rows be shuffled and this facility is available in BrainMaker. Therefore, another network was developed and experimented with after shuffling the facts. The test results obtained from this training showed an accuracy of 72% and 70% at testing tolerances of 0.4 and 0.2 respectively.

The next option used by the researcher to improve the performance of the previous networks was to test if a better network could be obtained by excluding one of the input variables in turn and training different networks. The exclusion of some of the variables resulted in a less performing networks, which indicates the importance of the excluded variable. For some variables, their exclusion resulted in a better performing network, which means that these variables are not important. A summary of the precision rate observed for the different variables at different testing tolerances is provided below.

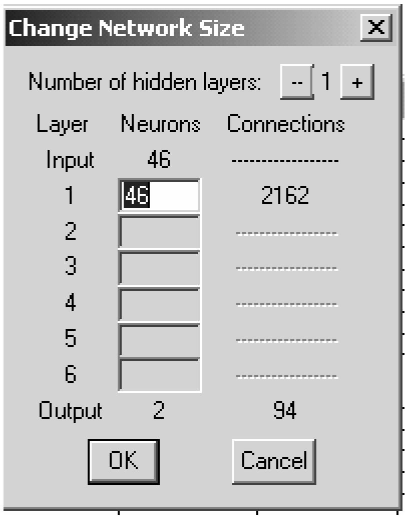
Table 13: List of variables excluded and the performance of those networks

Excluded Variable	***Testing Tolerance***			
	0.4	***0.3***	***0.2***	***0.1***
PA	73%	73%	71%	69%
ENVIRN	78%	76%	72%	67%
SEX	785	74%	73%	73%
AGE	70%	65%	62%	59%
HHRELIG	79%	74%	72%	66%
HHETHNIC	74%	69%	67%	63%
HHLITERAC	77%	74%	74%	68%
HHMEMBAVE	77%	74%	73%	72%
HHHEALTH	74%	70%	67%	63%
INMIG	83%	82%	76%	74%
OUTMIG	79%	72%	65%	62%
REL	75%	74%	73%	73%
HHWATER	76%	72%	68%	62%
HHROOF	75%	74%	71%	67%
HHLIVEST	79%	78%	74%	71%
WINDOWS	77%	73%	72%	65%

From the results presented in the above table, it is observed that the exclusion of the variables “ENVIRN”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, “HHROOF”, and “WINDOWS” resulted in a poor performing networks, which indicates the importance of these variables. Better performing networks were obtained when “REL”, “SEX”, “PA”, “INMIG”, “HHLIVEST” and “HHMEMBAVE” were excluded.

Therefore, for further trials of model building, the researcher decided to concentrate more on those important variables. By varying the composition of these variables, several networks were built to assess if a better network could be obtained. For example, three models were developed using the variables that were identified to have been important, but by varying the parameters used to build those models.

The first network developed in such a way had the following 14 input attributes: “PA”, “ENVIRN”, “SEX”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, “ HHMEMBAVE”, “HHROOF”, “HHLIVESTOK”, AND “WINDOWS”. This network had the following network size.



Number of hidden layers: 1

Layer	Neurons	Connections
Input	46	-----------------
1	46	2162
2		-----------------
3		-----------------
4		-----------------
5		-----------------
6		-----------------
Output	2	94

The network was trained using the following parameters:

Training Tolerance	***0.3***
Learning Rate	***0.6***
Smoothing Factor	***0.9***

During training, learning rate was gradually increased to 1. The model developed with the above parameters and input variables did not result in a better performing network. This network tested with the accuracy of 73%(i.e. it classified 80 of the of the 110 test cases correctly) with a testing tolerance of 0.4. With a tighter testing tolerance of 0.2 and 0.1the network tested with the accuracy of 65% (it classified 72 0f the 110 test cases correctly).

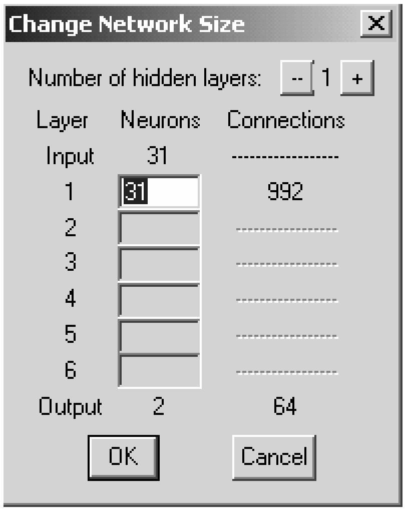
Another network was trained by using the parameters used to build the previous network, but by using the following 12 input variables: “PA”, “ENVIRN”, “SEX”, “AGE”, “OUTMIG”, “HHRELIGION”, “HHETHNICITY”, “HHLITERACY”, “HHHEALHTH”, “HHWATER”, “HHROOF”, AND “WINDOWS”. The variables excluded for this trial were: “RELATION”, “INMIG”, “HHMEMBAVE”, and “HHLIVEST”. The parameters were then slightly varied and tested. For instance training tolerance was reduced to 0.1 while the other parameters remained the same as in the above network. At a testing tolerance of 0.4, the accuracy of this network was 76%(i.e. it classified 84 of the 110 test cases correct). And with a testing tolerance of 0.2, the network tested with the accuracy of 73% (80 of the 110 test cases were classified correct). This result indicated that the exclusion of some attributes that were irrelevant for the learning task had an impact on the performance of the models built.

Numerous other experiments were also carried out by varying the parameters on the basis of suggestions made by different workers in the area and by the vendors of BrainMaker software. Form all above experiments, it became clear that varying of the parameters was not resulting in a significant change to the performance of the network. Rather, improvement in the performance of the network was observed when the composition of the input variables change.

Therefore, the researcher decided to proceed the experiment by using the default parameters which are strongly recommended by BrainMaker software. So, for the following experiments more emphasis was given on varying the composition of the input variables and shuffling of records than varying of parameters. The software providers also put that the default parameters are adequate for most problems (California Scientific Software, 1998).

The first network built by using the default parameters was trained with the 13 input variables namely “PA”, “ENVIRN”, “SEX”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALTH”, “HHWATER”, “HHROOF”, “HHLIVESTOK”, AND “WINDOWS”. The variables that were excluded for this trial were: “RELATION”, “INMIG”, and “HHMEMBAVE”. This network had 45 input neurons, 45 hidden neurons, and 2 output neurons. The accuracy rate of this network was 89% (it classified 98 of the 110 test cases correctly) at testing tolerance of 0.4 and it was tested with accuracy of 84%(it classified 93 of the 110 test cases correct) at a testing tolerance of 0.2. At tighter testing tolerance of 0.1, the network tested with accuracy of 80%(it classified 88 of the 110 test cases correct).

Another network was trained by using the following 10 input variables: “ENVIRN”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, “HHROOF”, “WINDOWS”. In this trial the variables excluded were: “PA”, “SEX”, “RELATION”, “INMIG”, “HHMEMBAVE”, and “HHLIVESTOK”. This network was trained using the default parameters. This network had the following network size:



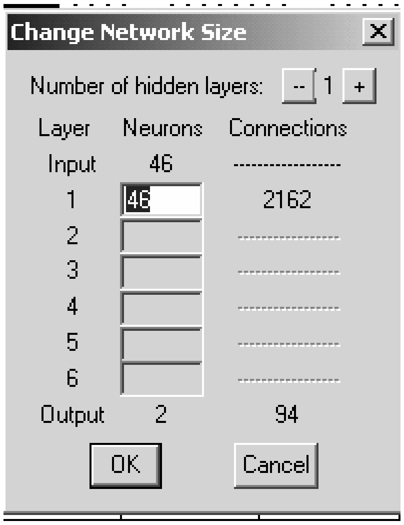
Change Network Size

Number of hidden layers: 1

Layer	Neurons	Connections
Input	31	----------------
1	31	992
2		----------------
3		----------------
4		----------------
5		----------------
6		----------------
Output	2	64

The test result of this network showed 88% accuracy (97 of the 110 test cases were classified correct) at testing tolerance of 0.4 and 80% (88 of the 110 test cases were classified correct) at testing tolerances of 0.2 and 0.1. Since there is some variation in the performance of the above two networks (models), the researcher continued the experiment by varying the combination of variables in an attempt to get a better performing network.

The next network was trained by using the following 14 input attributes: “PA”, “ENVIRN”, “SEX”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, “HHMEMBAVE”, “HHROOF”, “HHLIVESTOK”, AND “WINDOWS”. The variables excluded in this trial were: “REL”, and “INMIG”. This network was trained using the default parameters and has the following network size:



Change Network Size

Number of hidden layers: 1

Layer	Neurons	Connections
Input	46	-----------------
1	46	2162
2		-----------------
3		-----------------
4		-----------------
5		-----------------
6		-----------------
Output	2	94

This network showed an accuracy of 89% (i.e. it classified 98 of the 110 test cases correct) at testing tolerance of 0.4, and at testing tolerances of 0.2 and o.1 it was tested with accuracy of 85% (it classified 94 of the 110 test cases correct).

The next network trained was by using the default parameters and the following 9 input attributes: “ENVIRN”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, AND “WINDOWS”. This network had 29 input neurons and 29 hidden neurons. Using the above variables, several networks were trained and saved, and one of those networks achieved the best performing network. This network showed an accuracy of 93% (i. e it classified 102 of the 110 test cases correctly) at a testing tolerance of 0.4. To check the performance of this network on tighter testing tolerances, it was tested at testing tolerances of 0.2 and 0.1 and achieved a precision of 88% (i. e. it classified 97 of the 110 test cases correct).

As it was observed form the above experiments, several good performing networks were built. Particularly, the best performing network was observed when the following input variables were used: “ENVIRN”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, AND “WINDOWS”. This network seems to indicate that some attributes are more important to predict child mortality patterns than others.

For most of the networks saved, it was observed that there is a slight reduction in precision rate as testing tolerance became tighter.

The performance of the neural network models to predict mortality risks was also evaluated by using new cases that were not used in building of the models. For this purpose 10 records of children from both classes were selected as a running fact file to be used by the selected models. The main difference of these cases from the cases used for training and testing is that the output variable was not given in these new cases.

Then, the prediction accuracy of those models in classifying the cases correctly was evaluated. To determine the prediction accuracy a threshold value of 0.80 was set (i. e. only predictions made with a confidence level of 80% and above were considered valid). From all the models used to predict the outcome of those ten cases, one model (which was trained with 9 input attributes namely: ENVIRN", "AGE", "OUTMIG", "HHRELIG", "HHETHNIC", "HHLITERAC", "HHHEALHTH", "HHWATER", AND "WINDOWS") achieved an accuracy of 90%, where it classified 9 of the ten test cases correctly and only one case was misclassified.

The following table shows the results of the prediction analysis made by the above model.

Table 14: Results of the neural network prediction

OBSERVED	**PREDICTED**		Total
	Died	Alive	
Died	6	0	6
Alive	*1*	3	4
Total	*7*	3	10

As it can be observed from the above confusion matrix, the model wrongly misclassified one alive case as died.

The following table shows the sensitivity and specificity of the neural network model for each of the outcomes

Table 15: Sensitivity and Specificity of the neural network for the two classes

	Sensitivity	**Specificity**
Died	100%	75%
Alive	75%	86%

4.6 Models Built Using See5 Software

As it has been described at the beginning of this chapter, the researcher has experimented the application of See5 Decision Tree software by using the same dataset that has been used to build neural network models. The purpose of employing the decision tree method was to assess and evaluate the performance of the decision tree approach in building a predictive model as well as to see if this approach would result in better performing models to predict mortality patterns among children.

So, in this research project, the first trial of building See5 classifiers was carried out by using the default values provided by See5 software. When the default values are used, See5 constructs decision tree classifier. For this initial trial the researcher used 90% of the cases in the data file (i.e. testdata.data) for training and the rest 10% for testing purposes. For this trial, all the following 15 input attributes were used to build the classifier. The attributes were: “ peasant association”, “environment”, “age”, ‘inmigration”, “outmigration”, hhreligion”, “hhethnicity”, ‘hhliteracy”, “average meber of household”, “hhhealth”, “ hhwater’, “hhroof’, “:hhlivestock”, and “windows”. This initial attempt showed an accuracy of 91% on training data and 92 % on the test data. The following output shows evaluation of results achieved by the classifier on training data and test data.

Evaluation on training data (995 cases):

```
	    Decision Tree
	  ----------------
	  Size      Errors

	    52   89( 8.9%)   <<


	   (a)   (b)    <-classified as
	  ----  ----
	   462    52    (a): class Died
	    37   444    (b): class Alive
```

```
Evaluation on test data (111 cases):

            Decision Tree
          ----------------
          Size      Errors

            52    9( 8.1%)   <<


           (a)   (b)    <-classified as
          ----  ----
            45     7    (a): class Died
             2    57    (b): class Alive
```

An important feature of see5 is its mechanism to convert trees into collections of rules called rulesets. Invoking the Rulesets option causes rules to be derived from trees. Rulesets are generally easier to understand than trees since each rule describes a specific context associated with a class. Furthermore, a ruleset generated from a tree usually has fewer rules than the tree has leaves. Rules are also more accurate predictors of a cases class than decision trees. (http://www.rulequest.com). By default rules are ordered by class and sub-ordered by confidence. But, if "sort by utility" option is invoked, it provides an alternative ordering by contribution to predictive accuracy. Under this option, the rule that most reduces the error rate appears first and the rule that contributes least appears last.

Thus, to see the outcome of the ruleset option on the data prepared for the purpose of this research work, a classifier was constructed by invoking this option and by using the following variables: "ENVIRN", "AGE", "SEX", "OUTMIG", "HHRELIG", "HHETHNIC", "HHLITERAC", "HHHEALHTH", "HHWATER","HHMEMBAVE", "HHLIVESTOK", AND "WINDOWS". The classifier constructed by using these attributes had an accuracy of 91% on training cases (classified 903 of the 995 training facts correctly) and showed an accuracy of 92% (i.e. it classified 102 of 111 test cases correct) on test cases. Evaluation result of this classifier on training and test cases is presented as follows:

```
Evaluation on training data (995 cases):

                Rules
          ----------------
            No      Errors

            24   92( 9.2%)   <<


           (a)   (b)    <-classified as
          ----  ----
           467    50    (a): class Died
            42   436    (b): class Alive

Evaluation on test data (111 cases):

                Rules
          ----------------
            No      Errors

            24    9( 8.1%)   <<


           (a)   (b)    <-classified as
          ----  ----
            43     6    (a): class Died
             3    59    (b): class Alive
```

As it is observed from the evaluation results on test data, the total error rate of this classifier was 8.1% where it wrongly classified 6 Died cases as Alive and 3 Alive cases as Died.

A portion of the rules extracted by the above classifier is presented as follows:

```
See5 [Release 1.16]     Tue Jun 04 05:00:00 2002

    Options:
        Rule-based classifiers
        Use 90% of data for training

Class specified by attribute `status'

Read 995 cases (17 attributes) from testdata.data

Rules:

Rule 1: (71/2, lift 1.8)
        environ = L
        age <= 0
        hhliterac = Illitrate
        windows = N
        ->  class Died  [0.959]

Rule 2: (21, lift 1.8)
        hhethnic = Maraku
        hhhealth = SelfTreat
        ->  class Died  [0.957]

Rule 3: (23/1, lift 1.8)
        hhethnic = Maraku
        hhwater = River
       ➔  class Died  [0.920]

Rule 4: (211/18, lift 1.8)
        hhhealth = TradHeal
        ->  class Died  [0.911]

Rule 5: (14/1, lift 1.7)
        age <= 0
        sex = M
        hhethnic = Silti
        hhwater = River
        ->  class Died  [0.875]
       .

       .
Rule 11: (23, lift 2.0)
        sex = M
        hhrelig = Christian
        hhliterac = Literate
        hhmembave <= 6.2
        hhwater = Protwell
        ->  class Alive  [0.960]

Rule 12: (80/3, lift 1.9)
        hhhealth = Privateclinic
        hhwater = Protwell
        ->  class Alive  [0.951]

Rule 13: (75/3, lift 1.9)
        hhhealth = GovtHCU
        hhwater = Protwell
        ->  class Alive  [0.948]

Rule 14: (40/2, lift 1.9)
        age > 0
        hhrelig = Christian
        hhliterac = Literate
        hhwater = Pond
        ->  class Alive  [0.929]
```

As it is observed from the above rules, the classifier has used some attributes to construct rules and provided the class predicted by the rule. The numerical value, which appeared next to the predicted class, indicates the level of confidence of the predictor for the outcome or the predicted class.

As it is observed from rules 1 to 5 above, child mortality is associated with environment, household literacy, household health, age of the child, availability of windows in the house, household water, and even with household ethnicity.

> *For example, children who lived in lowlands and whose age is less than zero (between 0 to 11 months), and if their parents are illiterate, and if there is no window in the house, there is high probability of mortality for the child (See Rule1 above0.*

To determine the importance of the above rules and the attributes used to construct those rules, the association of the attributes with the predicted class predicted by rules were evaluated based up on comments given by domain experts and reports of previous research works.

As it is presented in rules 1 and 5,environmental (climatic) variations between highland and lowland communities revealed marked differences in child mortality. As per the discussions made with public health experts and pediatricians, the researcher proved that these variations in child mortality among highland and lowland communities seems appropriate since there is an epidemic of malaria and meningitis in rural lowlands than in highlands.

The classifiers constructed by using Ruleset have also revealed that lack of windows (i.e. poor housing conditions and crowding in the house) as a determinant factor for child mortality in the region. Particularly Infants (whose age is below 1 year) are more vulnerable for mortality due to lack of windows in the house. Similar findings have been observed in other studies conducted in the study area. For example, a study conducted by Desta (1994) indicated that out of 128 deceased infants, 101 lived in houses without a window. This study has also demonstrated that for infants, a fivefold ARI (Acute Respiratory Infection) mortality risk is associated with lack of a window.

From the above rules, it was also observed that the differences in access to, and utilization of, health care and preventive services were also identified as the major determinants of infant and child mortality in rural communities of the Butajira study area. As it can be seen from Rule 2 and Rule 4, if the major source of health care prevention for the child's parents is "self treat " or "Traditional

Healer", the child is more vulnerable for mortality than those whose parents major source of health care is " government health care unit" or "private clinic".

From rules1, 11, and 14 parental education was identified as a determinant factor for the survival of the child. So, based on those rules, it can be deduced that children whose parents are illiterate are more vulnerable to mortality than those whose parents are literate. The importance of this variable to determine the risk of child mortality has been reported in various studies. For example, a Nigerian study conducted by Adedoyin and Watts (1989) proved that, the risk of under-five mortality was threefold greater with non-educated parents. Another study by Desta (1994), which was conducted to determine under-five mortality and its public health determinants, revealed that, among parental factors, under-five mortality was significantly and independently associated with parental literacy. Other studies have also proved that parental literacy has been a strong predictor of infant and child survival.

An interesting finding identified from the above rules, particularly from Rules 2 and 3 is the association of child mortality and household ethnicity. As it is observed in those rules, children whose parents are from Silti and Maraku ethnicity have high mortality risks than children from other ethnic origins. For this association, the experts stated that such differences in child mortality by ethnicity or religion might be related to culturally determined differences in health behavior.

The source of water for the child's parents was also identified as a determinant factor for survival of the child. Particularly, as it is observed from rules 3, 5, 11, and 12 if the source of water for the child's family is river, they are more vulnerable to mortality than those whose source of water is protected well or pond.

As it can be observed from the previous discussions, the ruleset option had derived interesting and useful rules that can be applied to predict the survival probability of infants and children in the region.

The next option experimented by the researcher was to construct classifiers by using adaptive boosting option. Adaptive boosting is an important innovation incorporated in See5. The boost option with x trials instructs See5 to construct up to x classifiers. So selecting the Rulesets option and the Boost option with ten trials causes ten rulesets to be generated. The performance of the classifier constructed at each trial is summarized on a separate line, while the line labeled boost shows the result of voting all the classifiers.

In this research project, the potential of this option in building better classifier was also experimented and summary of the results obtained are presented as follows:

```
See5 [Release 1.16]     Thu Jun 06 00:33:48 2002

    Options:
        Rule-based classifiers
        10 boosting trials
        Use 90% of data for training

Class specified by attribute `status'

Read 995 cases (17 attributes) from testdata.data


Evaluation on training data (995 cases):

Trial       Rules
-----    ----------------
             No       Errors

   0         21    94( 9.4%)
   1         24   118(11.9%)
   2         30   114(11.5%)
   3         23   117(11.8%)
   4         37   104(10.5%)
   5         22   165(16.6%)
   6         24   111(11.2%)
   7         27   172(17.3%)
   8         24   138(13.9%)
   9         38   110(11.1%)
boost              53( 5.3%)   <<
```

```
               (a)   (b)    <-classified as
              ----  ----
               497    16    (a): class Died
                37   445    (b): class Alive


  Evaluation on test data (111 cases):

  Trial         Rules
  -----     ----------------
               No      Errors

     0         21    8( 7.2%)
     1         24   15(13.5%)
     2         30   13(11.7%)
     3         23   15(13.5%)
     4         37   11( 9.9%)
     5         22   20(18.0%)
     6         24   13(11.7%)
     7         27   26(23.4%)
     8         24   20(18.0%)
     9         38   11( 9.9%)
  boost              6( 5.4%)   <<


               (a)   (b)    <-classified as
              ----  ----
                52     1    (a): class Died
                 5    53    (b): class Alive
```

As it is observed from the above result, the boost option has resulted an accuracy of 95% on training cases (i.e. it classified 942 of the 995 training cases correct) and it achieved 95% accuracy on test cases (classified 105 of the 111 test cases correct). This classifier had wrongly classified one Died case as Alive and 5 Alive cases as Died from the total 111 test cases. The sensitivity and specificity of the above classifier on the test cases is presented in the following table.

Table 16: Sensitivity and specificity of decision trees for the two output classes

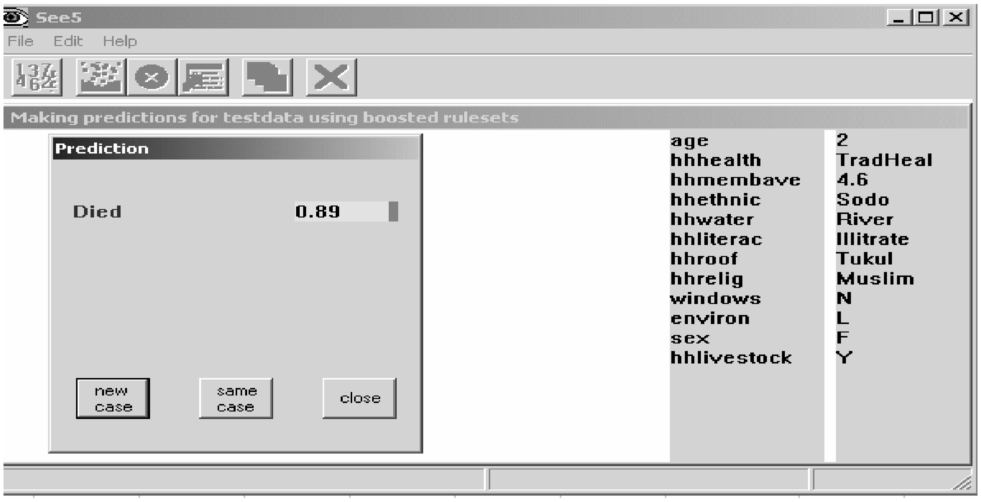
	Sensitivity	**Specificity**
Died	0.912	0.981
Alive	0.981	0.912

Then the classifier generated by using the adaptive boosting and ruleset options, was used to predict the class of new cases (i.e. the classifier was used to predict the survival probability of new cases). To do this, See5 provides a facility called "

Use classifier" which asks the values of the attributes needed to assign the new case to a class. After all the necessary attribute values have been entered, See5 shows the class to which the predicted case belongs. In this research work, this option was employed by the researcher to evaluate the accuracy of the constructed classifiers in predicting the survival probability of new cases (children).

To use the classifier to make predictions for new cases, the researcher used the ten records of children, which were prepared and used to evaluate the prediction performance of the neural network models. All the necessary values for the attributes requested by the classifier were supplied for each case, and the classifier predicted the survival probability of each new case based on those attributes. When the classifier makes predictions, it gives the degree of confidence in which the case belongs to the predicted class.

To determine the class of the child based on the degree of confidence provided by the classifier, a threshold value of 0. 80 was used by the researcher (i.e. predictions made with confidence level of 80% and above were considered valid). As a result, the classifier predicted the class of 9 of the 10 new cases correctly with in the specified threshold value. It misclassified only one of the Alive cases as Died. The following figure shows the screen of see5 while it makes predictions for one of the new cases that were used to evaluate the prediction performance of the classifier.



Making predictions for testdata using boosted rulesets

Prediction

Died 0.89

age	2
hhhealth	TradHeal
hhmembave	4.6
hhethnic	Sodo
hhwater	River
hhliterac	Illitrate
hhroof	Tukul
hhrelig	Muslim
windows	N
environ	L
sex	F
hhlivestock	Y

Figure 12: See5 screen while predicting new cases

As it observed from the above figure, the left pane of the screen shows the survival probability of the new case predicted by the classifier, and the right pane shows the attributes and their respective values for that case in which the classifier has used to predict the outcome of the new case.

The Overall model building process made by employing both neural network and decision tree techniques demonstrated that data mining is a method that should be considered to support public health care prevention and control activities at the district of Butajira. From the results obtained, it can be generalized that the survival probability of a child can be modeled simply by using public health determinants and risk factors gathered about children.

The model building process also revealed the importance of some variables that were not initially suspected to be very important. For example, variables like " household ethnicity", "age", and "availability of windows in the house" were identified as important variables to predict mortality patterns among children.
Although both neural network and decision trees showed comparable accuracy and performance in predicting the risk of child mortality, the decision tree approach seems more applicable and appropriate to the problem domain since it

provides additional features such as simple and easily understandable rules that are not available in neural network methods. Besides, even if it was not used in this research work, See5 software has an option call “ winnowing attribute” which can be easily employed to select the most relevant variables to construct classifiers or extract rulesets and to exclude irrelevant variables.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary and 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 application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as retail trade, health care, telecommunications, and banking. Specifically in the health care sector, data mining technology has been applied for patient survival analysis, prediction of prognosis and diagnosis, for outcomes measurement, to improve patient care and decision-making etc.

The objective of this research undertaking was to explore the possible application of data mining technology in the Ethiopian public health care context, and particularly at the district of Butajira, by developing a predictive model that could help health care providers to identify children at risk for certain ailments so that they can be treated before the condition escalates into something expensive and potentially fatal. Such a predictive model can then be applied in assisting child health care prevention and control activities in the region.

The methodology employed consisted of three basic steps; data collection, data preparation, and model building and testing. However, since a data mining task is an iterative process, these steps were not followed strictly in linear order. There were instances where there was a need to go back and forth between the different steps.

A data set totaling 1100 records of children was used to build and test both neural network and decision tree models. In order to build models that can predict the risk of child mortality, several models were built by employing both neural network and decision tree approaches. The best performing neural network model and decision tree classifier were then chosen and evaluated using ten previously unseen records of children.

Using the neural network approach, the best model was identified for the training made by using the default parameters (i. e. training tolerance of 0.1, learning rate of 1.0, and smoothing factor of 0.9) and the following 9 input variables: “ENVIRN”, “AGE”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, AND “WINDOWS”. This model had an accuracy rate of 93% (classified 102 of the 110 test cases correct) at a testing tolerance of 0.4 and was tested with accuracy of 88 % (classified 97 of the 110 test cases correct) at testing tolerances of 0.2 and 0.1.

Following the process of training and building neural network models, particularly until the overall prediction error reaches some acceptable minimum, ten new cases were presented to some best performing networks (models), and the accuracy of the selected models to predict the outcome of the new cases was evaluated. As a result, the model which was trained by using the default parameters and the above 9 input variables predicted the class of the 9 new cases correctly, and it misclassified one Alive case as Died.

Several classifiers were also constructed by using See5 decision tree software. From those classifiers, the best classifier was achieved when the ruleset and adaptive boosting options were used. The classifier was built by using the following attributes: “ENVIRN”, “AGE”, “SEX”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”,”HHMEMBAVE”, “HHLIVESTOK”, AND “WINDOWS”. This classifier resulted with an accuracy of 95% (i.e. it classified 942 of the 995 training cases correct) on training cases and it achieved 95% accuracy (classified 105 of the 111 test cases correct) on test cases.

The prediction performance of the See5 classifier was evaluated by using the ten unseen cases that were used to evaluate neural network models. As a result, the classifier predicted the outcome of the 9 cases correctly, where as it wrongly predicted one Alive case as Died.

During the course of model building, several important findings were observed. For example, the various trials made using both neural network and decision tree approaches revealed that some of the variables were consistently observed to be important. These variables were: *“ENVIRN”, “AGE”, “SEX”, “OUTMIG”, “HHRELIG”, “HHETHNIC”, “HHLITERAC”, “HHHEALHTH”, “HHWATER”, ”HHMEMBAVE”, “HHLIVESTOK”, AND “WINDOWS”.*

During the model building and testing process, the importance of the suggestions and opinions given by domain experts was also observed.

In general, encouraging results were obtained by employing both neural networks and decision tree approaches. Although both neural network and decision trees showed comparable accuracy and performance in predicting the risk of child mortality, the decision tree approach seems more applicable and appropriate to the problem domain since it provides additional features such as simple and easily understandable rules that can be used by non-technical health care professionals as well as health care planners and policy makers.

The trees and rules provided by decision tree models can also help to explain the prediction of a given outcome. However, neural networks do not provide methods to help explain the prediction. In support of this, Timm (2002) stated that although neural networks are perfect predictors, because of their black-box techniques, neural networks are not accepted in medical practice due to legacy, ethics, and scientific doubts. Thus, the researcher concluded that the decision tree approach is the method of choice to predict child mortality patterns using the BRHP epidemiological dataset, given its classificatory performance, rule extraction, and simplicity to understand and use.

The results obtained in this research work have proved the potential applicability of data mining technology to predict child mortality patterns based solely on demographic, parental, environmental, and epidemiological factors. The encouraging results obtained from both neural networks and decision trees indicate that data mining is really a technology that should be considered to support child health care prevention and control activities at the district of Butajira in particular, and at a national level in general. Specifically, such models could be used in settings outside of the hospital to support primary health care prevention activities and health service programs, which are aimed to reduce infant and child mortality.

5.2 Recommendations

This research work has uncovered the potential applicability of data mining technology to predict child mortality patterns based on demographic, socio-economical, parental, environmental, and epidemiological factors alone. Thus, based on the findings of this research work, the researcher would like to make the following recommendations particularly in relation to the possible application of data mining technology in supporting child health prevention and control activities aimed at reducing infant and child mortality in rural communities.

In the process of this research work, it was learnt that more research and development efforts need to be conducted to enable the full exploitation of data mining technology in the health care sector of this country. In particular, the following areas were identified as deserving further research work:

1. In this research work, an attempt has been made to assess the applicability of data mining technology to predict the likelihood of mortality for infants and children by using some set of variables that were considered important by experts. For a number of other variables, however, it remains to investigate further the effect of those variables to build models with better accuracy and performance than the models built in this research work.

2. It is indeed very important to assess the applicability of data mining techniques in predicting child mortality patterns by using clinical datasets gathered from different hospitals. This in turn would help to compare the results obtained with those clinical datasets to the results obtained using non-clinical or epidemiological datasets.

3. Since this study has used a small percentage of the ten years surveillance dataset to build neural network and decision tree models, it is appropriate to build more comprehensive models by using large training and testing datasets taken from the main database of the BRHP.

4. Although encouraging results were obtained in this research work, the researcher had experienced serious shortage of time to make some more trials to train, test and build more accurate and better performing networks or classifiers. The researcher has noted that the accuracy rate tends to increase to a desirable level with more and more trials of different combinations of the variables. It is therefore, recommended that more

time should be allocated to train and evaluate models with a desirable level of accuracy in prediction.

5. Although both the neural network and decision tree approaches resulted in an encouraging output, the application of other data mining techniques such as Belief Networks, which have also been proved to be important techniques in the health care sector should be experimented. Hence, it is recommended that other data mining techniques should also be tested to see if they could be more applicable to the problem domain.

6. More evaluation should be conducted particularly to assess whether rules qualify as a hypothesis for conventional research activities.

7. The possibility of incorporating the findings of this study in another (and operational) application should be explored.

REFERENCES

Adedoyin, M. A. and Watts, S. J. 1989. Child Health and Child Care in Okelele: an indigenous area of the city of Ilrorin, Nigeria. Social Science and Medicine, 29; 1333-1341.

Atiezenza, Felipe et.al. 2002. Risk stratification in Heart Failure using Artificial Neural Networks. Available URL : http://www.amia.org/pubs/sumpolia/D200367.pdf

Berry, Michael J. A. and Linoff, Gordon. 1997. Data Mining Techniques: for Marketing, Sales, and Customer support. New york; John Willy& Sons, Inc.

Bigus, Joseph p. 1996. *Data Mining with Neural Networks: Solving* Business Problems- from Application Development to Decision Support. McGraw-Hill: New York.

Brass, W. 1987. Problems in the Measurement of Child Mortality where statistical systems are limited. *Ann Soc Belg Med Trop*; 67: supp1: 57-70.

Bresnahan, Jennifer. 1997. Data Mining in the Health care : A Delicate operation. Available URL : http://www.Cio.com/archive/061597-mining-content.html

Cabena, P., et. al. 1998. Discovering Data Mining - From concept to Implementation, printice Hall, New Jersy.

California Scientific Software. 1998. URL: http://www.calsci.com/

California Scientific Software. 1998. BrainMaker, User's guide and reference Mannual.

Deogun, Jitender S. 2001. Data Mining: research Trends, Challenges, and Applications. Available URL: http://citeseer.nj.nec.com/deogun97data.html

Desta Shamebo. 1994. The Butajira Rural Health Project in Ethiopia: Epidemiological Surveillance for Research and Intervention in Primary Health Care. The Ethiopian Journal of Health Development , Vol. 8, Special Issue.

Desta shamebo. 1992. The Butajira Rural Health Project in Ethiopia: Epidemiological Surveillance for Research and Intervention in Primary Health Care. Scand Journal of Primary Health Care; 10: 389-96.

Dons, stephen M. and Wallace, Michael W. 2000. Mining Association Rules from a pediatric primary care Decision support system. Available URL: www.amia.org/pubs/symposia/D200658.PDF

Ekanem, E. E. 1985. Field epidemiology: Methodological constraints and limitations in the developing world. *Public Health*; 99: 33-36.

Fayyad, Usma, Piatetsky-shpiro, G. and smyth, padharic. 1996. From Data Mining to knowledge Discovery in Databases. Available URL: http://citeseer.nj.nec.com/fayyad96from.html

FDRE Central statistical Authority. 2001. The Year 200 Welfare Monitoring Survey. Addis Ababa, Ethiopia.

Fraser, Christopher M. 2000. Neural Networks: Literature Review from a Statistical Perspective. Hayward statistics. California State University, Hayward. Available URL: http://www.telecom.csuhayward.edu/~stat/Neural/CFPprojNN.htm

Frelichs, R. R. 1991. Epidemiologic Surveillance in Developing countries. *Annual review of Public Health*; 12: 257-80

Frohlich, Jochen. Neural Net Overview. 1999. Available URL: http://rfhs8012.fh-regensburg.de/~saj39122/jfroehl/diplom/e-1-text.html

Graetinger, Tim. 1999. Digging Up $$$with Data Mining – An executive guide, Discovery Corps, Inc. Availble URL: http://www.tdan.com/i010ht01.htm

Grove, Tom D. 2001. Neural Nets – Part I: Why are people more intelligent than machines? Available URL: http://umtii.fme.vutbr.cz/MECH/NN/tomgrl.html

Han, Jiawei and Kamber, Micheline. 2001. Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.

Knowledge Technology Inc. 2000. PC AI – Glossary of Terms. Available URL: http://www.primenet.com/pcai/New_Home_Page/glossary/pcai_glossary.html

Larvac, Nada. 1998. Data Mining in Medicine : Selected Techniques and Applications. Available URL.: http://citeseer.nj.nec.com/lavrac98data.html

Last, J. M. 1988. A dictionary of epidemiology, 2nd ed. New york: Oxford University Press.

Last, Mark and Kandel, Abraham. 2002. Automated Perceptions in Data Mining. Available URL : http://www.csee.usf.edu/~mlast/papers/perc_f1.pdf

Last, Mark, Maimon, oded, and Kandel Abraham. 2002. Knowledge Discovery in Mortality Records : An info-fuzzy Approach. Available URL: http://www.csee.usf.edu/softec/med_dm3.pdf

Lawrance, Jeannette. 1994. *Introduction to Neural Networks, Design, Theory, and Applications*. Nevada City: California Scientific Software Press.

Levin, Nissan and Zahavi, Jacob, 1999. Data Mining. Available URL: www.urbanscience.com/Data_Mining.pdf

Liu, H. and Motoda, H. 1998. Feature Selection for knowledge Discovery and Data Mining Available URL: http://www.databaseheadquarters.com/bookstore/management2/079238198XAMUS141630.shtml

Lloyd - Williams, Michael. 1997. Discovering the Hidden secrets in your Data - the data Mining approach to Information. Available URL: http://informationr.net/ir/3-2/paper36.html

Maimon, Oded, kandel, Abe and Last, Mark. 2002. Information Theoretic Fuzzy Approach to Knowledge Discovery in Databases. Available URL: http://www.csee.usf.edu/~mlast/papers/wsc_f2.pdf

Mannila, Heikki. 2002. Methods and Problems in data Mining. Available URL : http://www.cs.helsinki.fi/~mannila/

A Manual of the BRHP database: Unpublished material

Moon, Graham et. al. 2000. Epidemiology an introduction. Available URL : http://www.openup.co.uk.

Plate, Tony et. al. 1997. Visualizing the function computed by a Feedforward Neural Network. Available URL: http://pws.prserv.net/tap/papers/nc2000.pdf

Plate, Tony et.al. 1997. A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer. http://citeseer.nj.nec.com/plate96comparison.html

Prather, Jonathan C. *et. al*. 2001. Medical Data Mining : Knowledge Discovery in a clinical Data Woehouse. Available URL: http://www.amia.org/pubs/symposia/D004394.PDF

Pudi, Vikram. 2001. NeuralNetworks. Available URL: http://dsl.serc.iisc.ernet.in/~vikram/nn_intro.html

Raghavan, Vijay, Deogun, Jitender S. and Sover Mayri, 2002. Data Mining : Trends and Issues. Available URL: http://citeseer.nj.nec.com/138316.html

Raghavan, Vijav V., et al. 1998. A Perspective on Data Mining. *Journal of the American Society for Information Science*; 49(5): 397-402.

Rea, Allan. 20021. Data Mining : An introduction Student Notes. Available URL: http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html

Resnic, Fredric S. et. al. 2002. Development and evaluation of Models to predict Death and Myocardial Infraction Following Coronary Angioplasty and stenting. Harvard Medical school, Division of health sciences and technology, Boston: MA.

Rogers, Greg and Joyner, Elen. 2001. Mining your Data for Health care Quality improvement. Available URL : http://www.sinter.com.tw/SPSS

Rudolfer, Stephen M., Paliouras Georgios, Peers, and Ian S. 2002. A Comparison of Logistic Regression to Decision Tree induction in the Diagnosis of Carpal Tunnel Syndrome, Available URL: http://medg.lcs.mit.edu/ftp/wjl/cbr93/ml-paper.pdf.

Saarenvirta, Gray. 2001. Operation data Mining. Database programming and design Magazine. Available URL: http://www.db2mag.com/db_area/archives/2001/q2/saarenvirta.shtml

See5: An informal Tutorial. Available URL: http://www.rulequest.com/see5-win.html

Siganos, Dimitrois. 1997. Why Neural Networks. Available URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/ds12/article1.html

Skalk, David. 2001 Data Mining Blunders Exposed. Database programming and design Magazine. Available URL: http://www.db2mag.com/db_area/archives/2001/q2/miner.shtml

Stergiou, Christos, and Siganos, Dimitros. 1996. Neural Networks. Available URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vo14/cs11/reprt.html

Theeuwn, Marc; kappen, Bert; and Neist Jan. 2001. Neural Network analysis to predict Treatment outcome in patients with ovarian cancer. Available URL: http://www.mbfys.kun.nl/mbfys/people/bert/

Timm, Igno J. 1998. Automatic Generation of Risk classification for Decision support in critcal care. Available URL:

http://www.informatic.unti-bremen.de/~inti/paper/1998/gmds98-paper.pdf

Trybula, Walter J. 1997. Data Mining and Knowledge Discovery. *Annual Review of Information Science and Technology (ARIST)*; (32) : 197 - 229.

Two Crows Corporation. 1999. Introduction to Data Mining and knowledge discovery. Available URL: http://www.twocrows.com/

Waitman, Lamuel R. et. al. 1999. Knowledge Discovery in perioperative Databases using Rule induction : Hypothesis Testing, Decision support, and Clinical Guideline. Available URL: http://www.vuse.vanderblit.edu/~waitman/proposal/RussKDDproposalBody.htm

Wang, Samuel J. et. al. 2001. Using Patient-reportable clinical history factors to predict mycandidiat infraction. Available URL : http://www.elservier.com/local/compbiomed

Yemane Berhane, et al., 1999. Establishing an epidemiological field laboratory in rural areas- potentials for public health research and interventions. *The Ethiopian Journal of Health Development;* vol.134, special issue.

Z solutions. 1999. Neural Networks and data Mining. Available URL: http://www.zsolutions.com/sowhy.htm

ANNEXES

Annex 1: Sample dataset prepared for BrainMaker software

TYREF	PA	ENVIRN	REL	SEX	AGE	INMIG	OUTMIG	CAUSE	HHRELIG	HHEHNIC	HHLITER	HHMEMBAVE	HHHEALTH	HHWATER	HHROOF	HHLIVEST	Windows	DIED
10093	Meskan	H	3	F	4	N	N	NeoTet	Muslim	Meskan	Literate	7	TradHeal	Protwell	Tukul	Y	N	Y
10103	Meskan	H	9	F	1	N	N	Prematu	Muslim	Meskan	Illitrate	9	TradHeal	River	Tukul	Y	N	Y
10115	Meskan	H	3	M	2	N	N	Marasmus	Muslim	Meskan	Illitrate	5.5	GovtHCU	River	Tukul	Y	N	Y
10245	Meskan	H	9	F	1	N	N	Measels	Muslim	Meskan	Literate	8	Healthpost	River	Tukul	Y	N	Y
15489	Bati	L	9	F	1	N	N	ARI	Muslim	Meskan	Illitrate	4.6	TradHeal	Protwell	Tukul	Y	N	Y
15499	Bati	L	3	M	0	N	N	Measels	Muslim	Maraku	Literate	4.6	Healthpost	Protwell	Tukul	Y	N	Y
21124	Dobena	L	9	M	1	N	N	Measels	Muslim	Meskan	Literate	5	GovtHCU	River	Tukul	Y	N	Y
21153	Dobena	L	3	F	2	N	N	NeoTet	Muslim	Meskan	Illitrate	2.3	Pharmacy	River	Tukul	N	Y	Y
21162	Dobena	L	3	M	0	N	N	Accidents	Muslim	Meskan	Literate	4.6	GovtHCU	River	CorrIron	Y	N	Y
21163	Dobena	L	3	F	0	N	N	Kwash	Muslim	Meskan	Literate	4.6	TradHeal	River	CorrIron	Y	N	Y
21187	Dobena	L	3	F	5	N	N	ARI	Muslim	Meskan	Illitrate	5.3	TradHeal	River	Tukul	Y	N	Y
27781	Bido	H	9	M	1	N	N	ARI	Christian	Sodo	Illitrate	5	SelfTreat	River	Tukul	Y	N	Y
27782	Bido	H	9	M	0	N	N	Measels	Christian	Sodo	Illitrate	4.6	SelfTreat	River	Tukul	Y	N	Y
27882	Bido	H	3	M	1	N	N	ARI	Christian	Sodo	Literate	5	GovtHCU	River	Tukul	Y	N	Y
27887	Bido	H	9	F	3	N	N	Prematu	Christia	Sodo	Literate	5	SelfTreat	River	Tukul	Y	N	Y

TYREF	PA	ENVIRN	REL	SEX	AGE	INMIG	OUTMIG	CAUSE	HHRELIG	HHEHNIC	HHLITER	HHMEMBAVE	HHHEALTH	HHWATER	HHROOF	HHLIVEST	Windows	DIED
									n									
27888	Bido	H	9	F	0	N	N	Measels	Christian	Sodo	Literate	4.6	SelfTreat	River	Tukul	Y	N	Y
31658	Dirama	H	3	F	5	N	N	ARI	Muslim	Meskan	Literate	3.7	Healthpost	River	Tukul	Y	N	Y
31664	Dirama	H	3	M	3	N	N	Prematu	Muslim	Meskan	Literate	6	SelfTreat	River	Tukul	Y	N	Y
35749	Yeteker	H	9	M	0	N	N	Diarrhoaea	Muslim	Meskan	Illitrate	4.6	SelfTreat	Protwell	Tukul	N	Y	Y
35799	Yeteker	H	9	M	1	Y	N	Prematu	Muslim	Meskan	Literate	8	Pharmacy	River	Tukul	Y	N	Y
35814	Yeteker	H	9	F	2	Y	N	Others	Muslim	Meskan	Illitrate	5	Pharmacy	River	Tukul	Y	N	Y
35816	Yeteker	H	4	F	2	N	N	Others	Muslim	Meskan	Literate	3	SelfTreat	Protwell	Tukul	N	Y	Y
42435	Wrib	H	3	M	0	N	N	Others	Christian	Meskan	Illitrate	4.6	Healthpost	Protwell	Tukul	Y	N	Y
42438	Wrib	H	9	M	0	N	N	Diarrhoaea	Christian	Meskan	Illitrate	4.6	Healthpost	Protwell	Tukul	Y	N	Y
49533	Mjarda	L	3	M	2	N	N	ARI	Muslim	Maraku	Literate	7	Privateclinic	River	Tukul	Y	N	Y
49564	Mjarda	L	3	F	0	N	N	Diarrhoaea	Muslim	Maraku	Illitrate	4.6	Healthpost	River	Tukul	N	Y	Y
49592	Mjarda	L	3	M	4	N	N	Measels	Muslim	Maraku	Literate	2.8	Healthpost	River	Tukul	N	Y	Y
49608	Mjarda	L	9	M	1	N	N	Measels	Other	Maraku	Literate	6	Healthpost	River	Tukul	Y	N	Y
49618	Mjarda	L	9	F	2	N	N	NeoTet	Other	Maraku	Literate	7	Pharmacy	River	Tukul	Y	N	Y
53197	Hobe	L	9	M	7	N	N	Prematu	Muslim	Silti	Literate	4.6	TradHeal	Protwell	Tukul	N	Y	Y
53230	Hobe	L	3	M	1	N	N	ARI	Christian	Maraku	Literate	8	TradHeal	River	Tukul	Y	N	Y
53282	Hobe	L	3	F	0	N	N	Accidents	Muslim	Other	Literate	4.6	TradHeal	River	Tukul	Y	N	Y
53294	Hobe	L	3	F	1	N	N	Others	Muslim	Maraku	Literate	5	TradHeal	River	Tukul	N	Y	Y
11320	Meskan	H	3	F	9	N	N	N/A	Muslim	Meskan	Literate	5.9	Healthpost	River	Tukul	Y	N	N
11322	Meskan	H	9	F	7	N	N	N/A	Muslim	Meskan	Literate	6	Healthpost	River	Tukul	Y		N
11331	Meskan	H	3	M	9	N	N	N/A	Muslim	Meskan	Illitrate	6.2	Healthpost	Protwell	Tukul	Y	N	N

TYREF	PA	ENVIRN	REL	SEX	AGE	INMIG	OUTMIG	CAUSE	HHRELIG	HHEHNIC	HHLITER	HHMEMBAVE	HHHEALTH	HHWATER	HHROOF	HHLIVEST	Windows	DIED
11333	Meskan	H	3	M	2	N	N	N/A	Muslim	Meskan	Illitrate	7	Healthpost	Protwell	Tukul	Y	N	N
11335	Meskan	H	9	F	7	N	N	N/A	Muslim	Meskan	Illitrate	6.5	Healthpost	Protwell	Tukul	Y	N	N
15520	Bati	L	3	M	5	N	N	N/A	Christian	Maraku	Literate	9.6	Pharmacy	River	Tukul	Y	N	N
15523	Bati	L	9	M	3	N	N	N/A	Christian	Maraku	Literate	9	Pharmacy	River	Tukul	Y	N	N
21146	Dobena	L	3	M	5	N	N	N/A	Muslim	Meskan	Literate	7.8	Healthpost	Protwell	Tukul	Y	N	N
21147	Dobena	L	3	M	1	N	N	N/A	Muslim	Meskan	Literate	9	Healthpost	Protwell	Tukul	Y	N	N
21148	Dobena	L	4	F	5	N	N	N/A	Muslim	Meskan	Literate	7.8	Healthpost	Protwell	Tukul	Y	N	N
27806	Bido	H	9	M	7	N	N	N/A	Christian	Other	Illitrate	4.3	GovtHCU	Protwell	Tukul	Y	N	N
27811	Bido	H	3	F	2	N	N	N/A	Christian	Sodo	Literate	7	GovtHCU	Protwell	Tukul	Y	N	N
27812	Bido	H	3	F	5	N	N	N/A	Christian	Sodo	Literate	5.8	GovtHCU	Protwell	Tukul	Y	N	N
31661	Dirama	H	9	F	4	N	Y	N/A	Muslim	Meskan	Literate	4.5	Healthpost	Protwell	Tukul	Y	N	N
31684	Dirama	H	9	F	3	N	Y	N/A	Muslim	Meskan	Literate	11	GovtHCU	Protwell	Tukul	Y	N	N
35627	Yeteker	H	9	F	2	N	N	N/A	Muslim	Meskan	Illitrate	11	Pharmacy	Protwell	Tukul	Y	N	N
35633	Yeteker	H	3	M	0	N	N	N/A	Muslim	Meskan	Illitrate	4.6	Healthpost	Protwell	Tukul	N	Y	N
35636	Yeteker	H	9	F	3	Y	N	N/A	Muslim	Meskan	Illitrate	3	Healthpost	Protwell	Tukul	N	Y	N
35642	Yeteker	H	3	M	7	N	Y	N/A	Muslim	Meskan	Literate	4.3	Healthpost	Protwell	Tukul	Y	N	N
35643	Yeteker	H	3	F	5	N	Y	N/A	Muslim	Meskan	Literate	4	Healthpost	Protwell	Tukul	Y	N	N
42216	Wrib	H	3	M	3	N	N	N/A	Muslim	Silti	Literate	4	Healthpost	Protwell	Tukul	Y	N	N
42223	Wrib	H	3	M	1	N	N	N/A	Muslim	Silti	Illitrate	8	SelfTreat	Protwell	Tukul	Y	N	N
42226	Wrib	H	3	F	9	N	N	N/A	Muslim	Silti	Illitrate	4.2	SelfTreat	Protwell	Tukul	Y	N	N
42227	Wrib	H	3	F	7	N	N	N/A	Muslim	Silti	Illitrate	4.9	SelfTreat	Protwell	Tukul	Y	N	N
49386	Mjarda	L	3	M	7	N	N	N/A	Muslim	Maraku	Literate	8.5	TradHeal	Protwell	Tukul	N	Y	N
49387	Mjarda	L	3	F	5	N	Y	N/A	Muslim	Maraku	Literate	9.5	TradHeal	Protwell	Tukul	N	Y	N

Annex2: The Names File (testdat.names) created for See5 Application

```
Status|the class attribute.

pa:   Meskan, Bati, Dobena, Bido, Dirama, Yeteker, Wrib, Mjarda, Hobe, Buta04.
environ:    H, L, U.
age: continuous.
sex:    M, F.
inmig:  Y, N.
outmig: Y, N.
hhrelig:   Christian, Muslim, Other.
hhethnic:   Sodo, Dobi,Meskan, Maraku, Silti, Amhara, Oromo, Other.
hhliterac:    Literate, Illitrate.
hhmembave:   continuous.
hhhealth:    GovtHCU, CHA, Pharmacy, TradHeal, SelfTreat, Privateclinic, Healthpost, Nothing, Other.
hhwater:     River, Protwell, Unprotwell, Lake, Pond, Pipe, Other.
hhroof:       Tukul, CorrIron.
hhlivestock:  Y, N.
windows:  Y,N.
status:  Died, Alive.
TYREF:  label.
```

Annex3: Sample data set prepared for See5

```
Meskan, H, 4, F, N,N, Muslim, Meskan, Literate, 7, TradHeal, Protwell, Tukul, Y, N, Died,10093.
Meskan, H, 1, F,N,N, Muslim, Meskan, Illitrate, 9, TradHeal, River, Tukul, Y, N, Died, 10103.
Meskan, H, 2, M, N, N,  Muslim, Meskan, Illitrate, 5.5, GovtHCU, River, Tukul, Y, N, Died, 10115.
Meskan, H, 1, F, N, N,  Muslim, Meskan, Literate, 8, Healthpost, River, Tukul, Y, N, Died, 10245.
Meskan, H, 0, F, N, N,  Muslim, Meskan, Illitrate, ?, TradHeal, Protwell, CorrIron, Y, Y, Died, 10254.
Bati, L, 1, M, N, N,  Muslim, Meskan, Illitrate, 10, TradHeal, Protwell, CorrIron, N, Y,Died, 16612|, , ,
Bati, L, 0, F, N, N,  Muslim, Meskan, Literate, ?, Healthpost, Protwell, CorrIron, N,Y, Died, 16613|, , ,
Bati, L, 0, F, N, N,  Muslim, Meskan, Illitrate, ?, Healthpost, Protwell, Tukul, Y,N, Died, 16632|, , ,
Bati, L, 0, M, N, N,  Muslim, Meskan, Literate, 4.6, Healthpost, Protwell, Tukul, Y,N, Died, 16633|, , ,
Bati, L, 1, F, N, N,  Muslim, Meskan, Illitrate, 6, Healthpost, Protwell, Tukul, Y,N, Died, 16668|, , ,
Dobena, L, 0, M, N, N,  Muslim, Meskan, Illitrate, 9, SelfTreat, River, Tukul, Y,N, Died, 22347|, , ,
Dobena, L, 0, M, N, N,  Muslim, Meskan, Literate, 8, SelfTreat, River, Tukul, Y,N, Died, 22351|, , ,
Dobena, L, 0, M, N, N,  Christian, Silti, Illitrate, 4, Healthpost, River, Tukul, Y,N, Died, 22375|, , ,
Dobena, L, 0, M, N, N,  Christian, Silti, Illitrate, 3, Healthpost, Protwell, CorrIron, N,Y, Died, 22434|,
, ,
Dobena, L, 1, F, N, N,  Muslim, Meskan, Illitrate, 6.5, TradHeal, Protwell, Tukul, Y,N, Died, 22446|, , ,
Dobena, L, 3, F, N, N,  Christian, Silti, Literate, 3, Healthpost, Protwell, Tukul, Y,N, Died, 22541|, , ,
Dobena, L, 0, M, N, N,  Christian, Silti, Illitrate, 8, TradHeal, Protwell, Tukul, Y,N, Died, 22544|, , ,
Yeteker, H, 5, M, N, N,  Muslim, Meskan, Literate, 9.5, GovtHCU, River, Tukul, Y,N, Died, 35829|, , ,
Yeteker, H, 0, M, N, N,  Muslim, Meskan, Illitrate, 7, Healthpost, River, Tukul, Y,N, Died, 35874|, , ,
Yeteker, H, 3, M, N, N,  Muslim, Meskan, Illitrate, 7.3, TradHeal, River, Tukul, Y,N, Died, 35879|, , ,
Yeteker, H, 1, M, N, N,  Muslim, Meskan, Illitrate, 8, TradHeal, River, Tukul, Y,N, Died, 35881|, , ,
Yeteker, H, 0, M, N, N,  Muslim, Meskan, Illitrate, 6.7, Healthpost, Protwell, Tukul, Y,N, Died, 35971|, ,
,
Yeteker, H, 0, M, N, N,  Muslim, Meskan, Illitrate, 6, TradHeal, River, Tukul, Y,N, Died, 35973|, , ,
Mjarda, L, 5, F, N, N,  Muslim, Silti, Literate, 6.4, Healthpost, Protwell, Tukul, Y,N, Alive, 49411|, , ,
Mjarda, L, 1, F, N, N, Muslim, Silti, Illitrate, 3, Healthpost, Protwell, Tukul, Y,Y, Alive, 49415|, , ,
Mjarda, L, 0, M, N, N,  Muslim, Silti, Literate, 4.6, Healthpost, Protwell, Tukul, Y,N, Alive, 49419|, , ,
Mjarda, L, 2, M, N, N,  Muslim, Silti, Literate, 3, Healthpost, Protwell, Tukul, N,Y, Alive, 49422|, , ,
Mjarda, L, 0, M, N, N,  Muslim, Silti, Literate, 4.6, Healthpost, Protwell, Tukul, N,Y, Alive, 49423|, , ,
Mjarda, L, 2, M, N, N,  Muslim, Maraku, Literate, 5, GovtHCU, Protwell, Tukul, N,Y, Alive, 49434|, , ,
Dobena, L, 4, M, N, N,  Muslim, Other, Literate, 6.1, Healthpost, Protwell, Tukul, Y,N, Alive, 21548|, , ,
Dobena, L, 4, F, N, N,  Muslim, Other, Literate, 6.3, Healthpost, Protwell, Tukul, Y,N, Alive, 21550|, , ,
Dobena, L, 2, F, N, Y, Christian, Other, Literate, 7, Healthpost, Protwell, Tukul, Y,N, Alive, 21561|, , ,
Dobena, L, 0, F, N, N,  Christian, Other, Literate, 7.6, Healthpost, Protwell, Tukul, Y,N, Alive, 21563|, ,
,
Dobena, L, 5, M, N, N,  Muslim, Meskan, Literate, 4.4, Pharmacy, Protwell, Tukul, Y,N, Alive, 21571|, , ,
Buta04, U, 5, F, N, N,  Christian, Meskan, Literate, 7.6, Healthpost, Pond, CorrIron, N,Y, Alive, 59023.
Buta04, U, 3, F, Y, N,  Christian, Meskan, Literate, 4.6, Healthpost, Pond, CorrIron, N,Y, Alive, 59024.
Buta04, U, 1, F, N, N,  Christian, Other, Literate, 5, Healthpost, Pond, CorrIron, N,Y, Alive, 59048.
Buta04, U, 2, F, Y, N,  Christian, Other, Literate, 5, Healthpost, Pond, CorrIron, N,Y, Alive, 59052.
Buta04, U, 1, F, N, N,  Christian, Silti, Literate, 4, Healthpost, Pond, CorrIron, N,Y, Alive, 59057.
Buta04, U, 4, F, Y, N, Christian, Silti, Literate, 2.8, Healthpost, Pond, CorrIron, N,Y, Alive, 59059.
Buta04, U, 2, F, N, Y,  Christian, Silti, Literate, 4, Healthpost, Pond, CorrIron, N,Y, Alive, 59060.
Buta04, U, 5, F, N, Y, Christian, Other, Literate, 3.3, GovtHCU, Pond, CorrIron, N,Y, Alive, 59064.
```