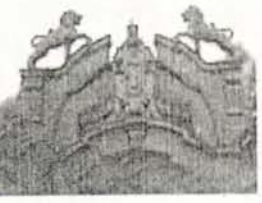


ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE
AND
SCHOOL OF PUBLIC HEALTH

MINING VITAL STATISTICS DATA: THE CASE OF BUTAJIRA RURAL HEALTH PROGRAM

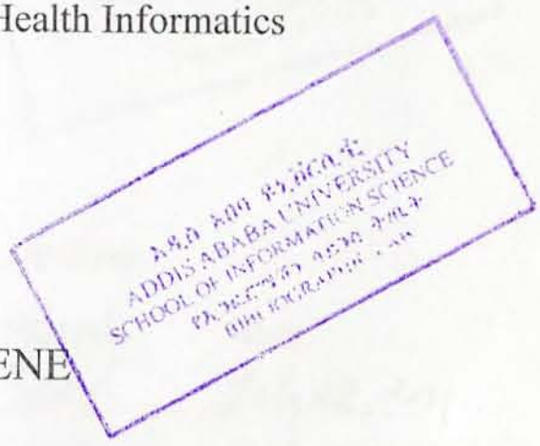
TADESSE BEYENE

JUNE 2011

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

MINING VITAL STATISTICS DATA: THE CASE OF BUTAJIRA RURAL HEALTH PROGRAM

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



By

TADESSE BEYENE

JUNE 2011

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

MINING VITAL STATISTICS DATA: THE CASE OF BUTAJIRA RURAL HEALTH PROGRAM

By

TADESSE BEYENE

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Mahder Alemayehu	Chairperson		July 28, 2011
Million M.	Advisor(s),		July 28, 2011
Mitike Molla	Advisor(s),		July 29, 2011
Worksheet Lamenew	Examiner,		July 28, 2011

Declaration

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

28/07/11
Date

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

Advisor

Advisor

DEDICATION

This thesis is dedicated to My Brother, Mr. Hailemariam Beyene.

ACKNOWLEDGEMENT

First of all I would like to thank God for his blessing and being always besides me during my pleasure and trouble times. My deepest gratitude, then, goes to Dr. Million Meshesha and Dr. Mitike Molla for their guidance, patience, support and encouragement throughout my study at the Addis Ababa University, which led to this thesis.

I am extending my sincere thanks to the database manager, Dr. Alemayehu Worku, of Butajira Rural Health Program (BRHP) for his assistance in availing and explaining the nature of the BRHP data and commenting on the knowledge obtained from the research findings.

Many individuals have contributed support, idea expertise, insight and time to make this study possible. With this regard I would like to thank my lovely wife, Mrs. Yalemhiwot Tariku, for the mental peace and love she gives me in doing this research.

My sincere appreciation goes to thank my parents, brothers, sisters, sister-in-law and especially my mother, Mrs. Tsehaynesh Asfaw for their undying prayers, love, encouragement and moral support. 'Thank you'

Last but not least I want to thank all my friends, especially Fitsum Nigussie and Abiy Zemede and colleagues who stayed by me throughout this period of time constantly encouraging me to work hard and work at the Addis Ababa University a very pleasurable one. I am evenly thankful to all others who assisted me in doing this study.

Table of contents

Contents **Pages**

ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
API	Application Program Interface
ARI	Acute Respiratory Infection
ARRF	Attribute Relation File Format
BRHP	Butajira Rural Health Program
CART	Classification and Regression Tree
CRISP-DM	CRoss Industry Standard Process for Data Mining
CSV	Comma Separated Value
DM	Data Mining
DME	Data Mining Engine
DSA	Demographic Surveillance Area
DSS	Demographic Surveillance Site
EBM	Evidence Based Medicine
EM	Expectation Maximization
EPI-info	Epidemiological Information
FPGrowth	Frequent Pattern Growth
GIS	Geography Information Systems
HI	Health Informatics
HIS	Health Information System
HIV	Human Immunodeficiency Virus
ID	Identification
IT	Information Technology
JDM	Java Data Mining
KDD	Knowledge Discovery in Database
KDLC	Knowledge Discovery Life Cycle
MIS	Management Information System
MOR	Mining Object Repository
NB	Naive Bayes
NN	Neutral Network
OLAP	Online Analytical Process

REC	Research and Ethics Committee
ROC	Receiver Operating Characteristics
SAS	Statistical Analysis Systems
SMOTE	Synthetic Minority Oversampling Technique
SNB	Simple Naive Bayes
SNNPRS	Southern Nations, Nationalities and Peoples Regional State
SPSS	Statistical Package for Social Sciences
SQL	Structured Query Language
UN	United Nations
USA	United State of America
VA	Verbal Autopsies
VCT	Voluntary Counselling and Testing
VSD	Vital Statistics Data
WEKA	Waikato Environment for Knowledge Analysis

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Data mining is a relatively new field whose major objective is to extract knowledge hidden in large amounts of data. Vital statistics data offer a fertile ground for data mining by providing a valuable source of information regarding the health status of a population. One of the most important public health functions is monitoring of a population's health status. At all levels of the health delivery structure a well organized health information system is vital for identifying the health needs of populations and for planning, implementation and monitoring of health interventions.

The aim of this study is to discover knowledge that can be used to gain insight into various aspects of mortality in the selected rural area of the country. The study explores the death aspect of the vital statistics data in the Butajira Rural Health Program- BRHP database at Butajira, Ethiopia.

A data mining tool called Weka is used to build predictive model of 95,220 cases over an eighteen-year period. A historical cohort study analysis of vital statistics is conducted. It follows a JDM process modeling. This study apply classification algorithm, such as to extract interesting knowledge from temporal data on BRHP database.

The results obtained in the study contain valuable new information. These results conveyed some interesting findings. The classification algorithm reveals that the result indicates for the BRHP dataset, over 90% accurate results are possible for developing classification rules that can be used in prediction.

From this result the researcher concludes that the vital statistics data can help to predict using the application of data mining classification technique given the limitation of this study. In general, the result from this study is encouraging.

Keywords: vital statistics data, Machine Learning, data mining, predictive models, classification, Weka.

CHAPTER ONE

INTRODUCTION

1.1. Background

The healthcare industry is among the most information intensive industries. Health information, knowledge and data generated globally are growing up. It has been estimated that an acute care hospital may generate five terabytes of data a year [1]. The ability to use these data extract useful information for quality healthcare is crucial.

Health informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place. It is known that "Discovery of HIV infection and Hepatitis type C were inspired by analysis of clinical courses unexpected by experts on immunology and herpetology, respectively" [2]

.

With the rapid development of information technology, a variety of information has increased and there exists generally the phenomenon of "data rich but information poor" [3]. How to excavate the valuable knowledge and the information from the mass data is the main challenge information management faces these days. The challenge of extracting knowledge from data is of common interest to several fields, including statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing [4].

Data mining is about using statistical and machine learning techniques to find relationships in a set of (usually) large and complex data. It is a promising research area since it is a new emerging research field by gaining insight into various aspects of the accumulated data.

The Data Mining (DM) technique has become an established method for improving statistical tools to discover knowledge exist within the data which can be used to predict future trends [5]. The data mining process involves identifying an appropriate data set to "mine" or sift through to discover data content relationships. Data mining tools include techniques like case-based reasoning, cluster analysis, data visualization, fuzzy query and analysis, and neural networks. Data mining sometimes resembles the traditional scientific method of identifying a hypothesis and then testing it using an appropriate data set. Sometimes, however, data mining is reminiscent of what happens when data has been collected and no significant results were found and hence an ad hoc, exploratory analysis is conducted to find a significant relationship [2]. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place [5].

Data mining as an analytic process designed to explore large amounts of data in search for consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The process thus consists of three basic stages [6]: exploration, model building or pattern definition, and validation/verification.

What distinguishes data mining from conventional statistical data analysis is that data mining is usually done for the purpose of "secondary analysis" aimed at finding unsuspected relationships unrelated to the purposes for which the data were originally collected [2]. Such knowledge discovery mechanism is crucial for competitive advantage and sustainable development.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools. Data mining answers business questions that traditionally were too time-consuming to resolve. Data mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [2].

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem and regression analysis [7].

Classical statistics typically includes the following kinds of analyses [8]: **simple** (view one or more measures that can be sorted and totaled), **comparison** (view one measure and sort or total based on two or more dimensions), **trend** (view measure over time), **variance** (compare one measure at different times such as "sales" and "sales a year ago"), and **ranking** (top 10 or bottom 10 products sold). This enables users to drill down within a dimension to see more detailed data at various levels of aggregation. Users can also filter data, that is, focus their analysis on a subset of records in the database.

Structured query languages (SQL) are well known software tools with very little freedom for manipulations and SQL is useful for finding information, as long as the user knows perfectly what he or she is searching for [9]. Once the user provides the Query the processor will provide the user with the exact answer that is required for the solution. Sometimes, for instance, we come across cases where the patient has symptoms of fever and sweating. SQL cannot provide us with a diagnosis or decision about whether the patient is having a headache or a cold based on the information provided [9].

The proliferation, ubiquity and increasing power of computer technology has increased data collection, storage and manipulations. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing [7]. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms, decision trees and support vector machines. Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. It has been used for many years by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports [7].

Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery [9].

One of the most important public health functions is the monitoring of a population's health status. Vital statistics data provide a valuable source of information regarding the health status of a population and hence offers a fertile ground for data mining.

Vital event registration system, which is the continuous, permanent and compulsory recording of the occurrence and characteristic of vital events, is the basis for developing legal, administrative and statistical information system that protects and safeguards most rights and privileges of individuals (citizens) endorsed in the numerous conventions and recommendations of the United Nations (UN) [10]. However, it is nonexistent or only partially applied in many developing countries. Given the paucity of vital-events registration and knowledge on population or health-status trends in such settings, demographic and health surveys have been introduced for health planning, practice, evaluation, and allocation of resources. The most commonly used types of vital statistics data in public health are data on births and deaths [11].

The Butajira Rural Health Programme (BRHP) is one of the earliest Demographic Surveillance Site (DSS) that produces vital statistics data in Ethiopia. DSSs rely on regular community-based surveillance as a means of vital event registration. It was initiated in mid-1986 with a complete census of the 10 randomly sampled kebeles[1] in Meskan and Mareko [11]. Soon after, by January 1987, a DSS with continuous registration of vital events was initiated. The major aims were to develop and evaluate a system for continuous registration of births, deaths and migratory events to generate valid data on fertility and mortality and provide a study base for essential health research and intervention [11].

[1] Kebele is the smallest administrative unit in Ethiopia

The BRHP DSS is primarily a collaborative research project undertaken by the Department of Community Health (now school of public heaalth), Faculty of Medicine, Addis Ababa University, Ethiopia, and the Division of Epidemiology, Department of Public Health and Clinical Medicine, Umea University, Sweden. The collaboration started as a doctoral-study project. Later, it grew into a departmental collaboration and included the development of the study-base infrastructure and involvement of a multidisciplinary group of researchers. The original DSS population in 1987 was around 28,000 and grew over 10 years to about 37,000 active individuals, currently with more than 61,500 individuals involved at BRHP during April 2011 monitoring [5].

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 unfamiliar 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 the state-of-the-art technology in the Ethiopian health-care context. Thus, in this research work, the researcher is motivated to explore the potential applicability of data mining technology to gain insight mortality patterns in rural Ethiopia by using the BRHP database. For reasons of familiarity and availability of electronic data, the researcher chose the Butajira Rural Heath Program (BRHP) to conduct the study.

1.2. Statement of the Problem and Justifications

Research in data mining is exploring data to acquire knowledge and understanding that enable the development and implementation of technology-based solutions to heretofore unsolved and important business problems.

In any exploration (health related cases, criminal reasons or any other issue) an investigator needs to sort through multiple events and factors and decide which ones have a likely causal relationship on a specific outcome. Sorting among the known candidates to determine the actual culprit requires data collection, and when the collected data

olves complex multivariate relationships, data mining may provide insight for covery [12].

organizations, including many health centers, are increasingly creating or accessing ger and larger databases, and analysts looking at this data would be wise to learn more out what data mining can do for them. In normal statistics, one is trying to relate two ngs (say, smoking and cancer rates) and then try whether or not that link is true. In data ning, a researcher attempt to find relationships that we may not even know exist - re just hoping that if one look at enough information, certain links (e.g. that people o smoke a lot are more likely to get cancer) may pop out of it.

e issues in data mining are noisy data, missing values, static data, sparse data, dynamic a, relevance, interestingness, heterogeneity, algorithm efficiency, size and complexity data. These types of problems often occur in large amounts of data. Once these blems resolved and a possible relationship is found, we'll test them on other data to if they're true or not. This theory challenges researchers to create artifacts that enable anizations to overcome the acceptance problems predicted.

a for the demographic surveillance site (DSS) were initially entered as text strings, but BRHP DSS has, since 1994, used software based on the dBase IV platform. Data are rently entered in Butajira, which allows any inconsistent questionnaires to be nediately sent back to the field. This is a significant improvement over earlier ctice, which was to centralize data operations in Addis Ababa [11].

site manipulates and analyzes data with dBase, Epi-Info, and the Cohort program, eloped by Umeå University, which does person–year-based analyses of events in amic cohorts. National and international publications and scientific conferences have n the main routes to disseminate this information. Community feedback meetings e been held periodically [5].

DSS were able to collect data, sift through and analyze the mortality data during their es because the volume of information was manageable. Today, the size of the base, the amount of electronic data gathered containing massive amounts of

information which keeps on growing year after year; make it almost impossible to accomplish what the pioneers do.

This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector. Typical problems that data mining addresses are how to classify data, cluster data, find associations between data items, and perform time series analysis. Numerous data mining techniques have been invented for each type of problem. Each problem requires data mining techniques to analyze large quantities of data.

The problems with analyzing the adverse events in the BRHP include the following:

- The feedbacks reaching the program are most of the time spontaneous, not all problems gets to the ears of the BRHP.
- They can only internally compare with other cases from the same company, not with the cases from the competing programs.

Unfortunately, causes do not only exacerbate death, one might also get adverse events. In some situations, there might be multiple possible causes. For the program, the mandatory BRHP exist (Periodic Safety Update Report), but these are only listings and tables, no new knowledge discovery is done.

Numerous studies have been conducted on various aspects of health parameters and determinants by using the information gathered by the BRHP epidemiological surveillance system. Among the major research findings in the area of reproductive health was the observation of the high prevalence of domestic violence [13], which is a serious concern for women in the area affecting their social, psychological and physical well-being. Specifically, previous studies [11] conducted on mortality indicates the existence of very high rate of 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.

The problems of previous research efforts are not only related to the small proportion of the database used, but data analysis is also conducted by using simple statistical techniques (such as regression and verification techniques). As Anagaw [14] cited 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. Using data mining techniques, however, we can find out which adverse events occur more frequently with specific cases in order to gain insight into the whole recorded datasets.

This study, therefore, aims to apply the data mining approach and extract hidden patterns and knowledge related to death aspect of the vital statistics data in the BRHP which enhances decision making and effective management.

To this end, the study attempts to answer the following research questions:

- What DM initiatives are currently underway in the program and stakeholder group?
- Have there been coordinated, aligned, communitywide efforts, strategies, or plans related to DM across multiple stakeholders in the program?
- What opportunities have been seen for accelerating adoption of DM to improve quality analysis, such as linking disparate efforts, creating synergies across efforts, etc?

1.3. **Objective of the Research**

1.3.1. **General objective**

The general objective of this study is to explore suitable data mining technique and method for knowledge discovery that can be used for various aspects of demographic variables using BRHP database in order to predict outcomes for future situations with

health issues related to the death, to offer an aid to decision- or policy-making process, and to provide useful information services to the customers.

1.3.2. Specific objectives

To achieve the general objective, the specific objectives conducted in this research are the following:

- To conduct a thorough review of literature that can support the study in the area of applying data mining technology on health care in general and vital statistics in particular.
- To select appropriate data mining tool and technique (like classification, clustering, association, etc.) for performing vital statistics data mining function.
- To generate good quality datasets of vital statistics that can be used for data mining task
- To design suitable data mining method for creating predictive model using vital statistics data
- To evaluate data mining models and select the best model that is more appropriate to the problem domain.
- To test the selected model using the selected tool, technique and algorithm.
- To recommend further research directions for discovering knowledge from vital statistics data.

1.4. Scope and Limitation of the Research

The scope of the research is delimited to one of the rural health program at Ethiopia centered at Butajira DSS with the assumption of the database is a representative of the other kebeles of the other districts. Since data of each kebele of the selected district are reported on a monthly/quarterly basis to the central database located in Addis Ababa, the scope is therefore not confined to birth, new household, and migration data in the database rather it is a holistic representative of the BRHP.

The applicability of data mining to BRHP database is limited to develop and test the model instead of deploying the model at the Butajira district since the study is carrying out for academic achievement. That is, 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.

Although several techniques and algorithms are available in data mining tasks, we concentrate on applying predictive classification learning rules. Within each type of learning methods, decision tree is considered for classification vital statistics data.

Another limitation for this study is that only 95,220 of records are taken due to the scalable constraint of Weka heap size (memory size) as discussed in data selection and preparation procedure. Moreover, delay of approval of the proposal consumed much of the time that can have been used for actual research work. Such kind of constriction forces the researcher to be limited with only one technique and an algorithm instead to make further more different experiments for discovering knowledge.

1.5. **Research Methodology**

Knowledge of the methods for collecting or compiling data at the DSS sites is essential because these methods influence the ways that data are processed, analyzed, and interpreted. The most common demographic methods used in data collection are censuses, sample surveys, and vital-events registration systems. From these concepts the researcher tries to apply data mining research using the following methodologies and procedures.

The purpose of specify methodology is to provide insight with fundamental knowledge of data modeling and design; the tools and techniques of data analysis using data mining technology; to acquaint beneficiaries with data mining concepts, and techniques; and to prepare data for further analysis in database. Specifically this section provides an understanding of the concepts for knowledge discovery in database.

1.5.1. **Data source**

For any data mining task, the primary requirement is availability of data in any format [15]. In this study, secondary data source is used. This data is organized by collecting facts from households located at Butajira and surrounding areas. Hence this research is conducted over the available database; the source of data is the Butajira Rural Health Program. The 18 years separate vital statistics data is used as a base for data collection, which is kept for the study in the School of Public Health, Addis Ababa University. This separate eighteen years' data set contains a total of 236,549 records of individuals registered in all the ten villages of the BRHP study area. Since the data are very sensitive in nature, the privilege to view the data is allowed after proposal development.

Therefore no further data collection mechanisms are employed as the collected data is ample enough to undertake the planned research.

1.5.2. **Study design**

In this study, the researcher attempt to explore the data mining approach on the death aspect of the vital statistics data in the BRHP. For this purpose, a data mining tool called weka is used.

Historical cohort study is conducted on vital statistics using longitudinal database in Butajira Rural Health Program (BRHP) to discover knowledge that can be used to gain insight into various aspects of mortality in the selected rural area of the country, to predict outcomes for future situations with health issues related to the death, to offer an aid to decision- or policy-making process, and to provide useful information services to the customers.

In this study, the researcher follow the Java Data Mining (JDM) standard process by using historical cohort study on longitudinal database in Butajira Rural Health Program (BRHP) to discover knowledge that can be used to gain insight into various aspects of mortality. The data mining process is an iterative or cyclic process that involves a number of stages. As described in Java Data Mining (JDM) with details of process life cycle,

below Figure 1.1 shows the java data mining process that the researcher follow to mine the vital statistics data.

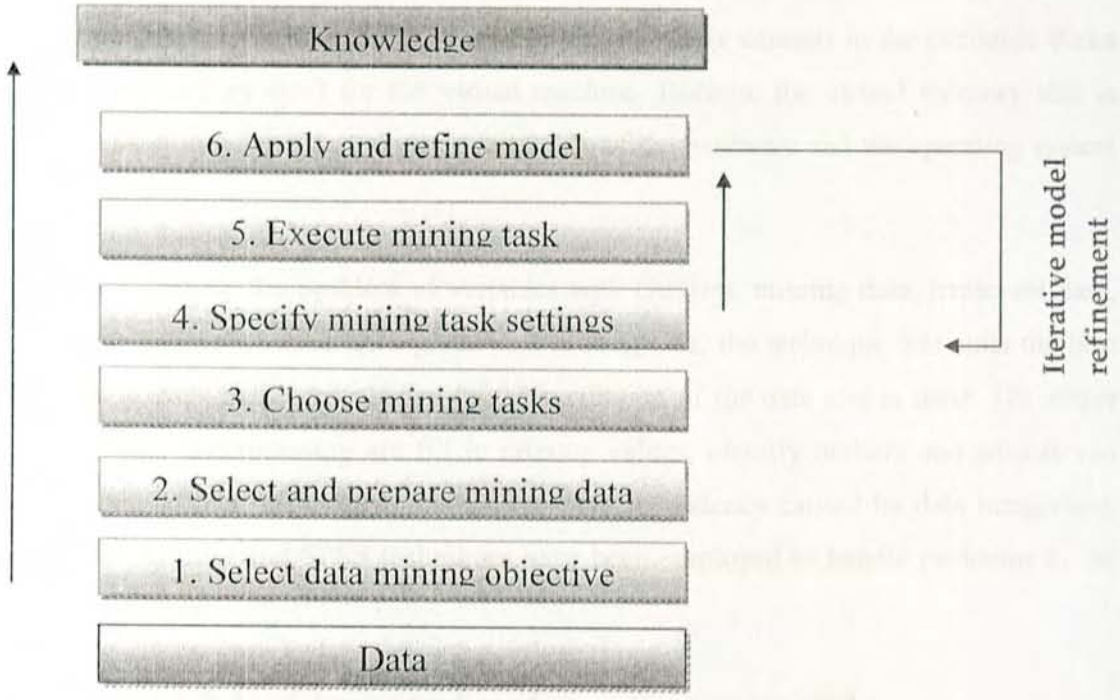


Figure 1.1 Java Data mining process

1.5.2.1. Select data mining objective

In this regard, the researcher makes efforts by reading different published articles and consulting the concerned people to understand the business settings and to gain knowledge for the problem under study as stated in detail at the next chapter. In line with this, a data mining research problem is planned with set of objectives to approach the problem under this chapter of the paper.

1.5.2.2. Select and prepare data

The list of sampled cases which are registered as demographic data is identified from the whole 18 years database of the BRHP. We use Weka unsupervised resample and oversampling techniques to sample 66,123 cases (95,220 cases after SMOTE) from the whole 236,549 cases in the dataset. In order to take the sample for this study, we first sort

the dataset according to the time period (in this case year) the data recorded and then we select a 25% of the cases from each year (i.e 18 years) without replacement technique available in Weka software. The reason that forced to select a sample for the experimentation is that the difficult for running the whole datasets in the available Weka heap size (memory size) for the virtual machine. Because the virtual memory size is system (machine) dependant on the availability of the hardware and the operating system loaded on it.

In order to handle the problem of variables with Outliers, missing data, irrelevant data, etc in the data set created for a given data mining task, the technique that suits the best and that doesn't make any change on the prediction of the data sets is used. The major tasks in data preprocessing are fill in missing values, identify outliers and smooth out noisy data, correct inconsistent data and resolve redundancy caused by data integration. These Weka filter and SPSS techniques have been employed to handle problems in the obtained dataset.

1.5.2.3. Choose and configure the mining tasks

The data mining task appropriate for the problem under consideration is selected to be classification. Having selected a mining task, we would then configure that task with parameters suitable for the task.

In this study, we identify the outcome with respect to the classification task, thus we concentrate mainly on supervised learning method in data mining.

1.5.2.4. Select and configure the mining algorithms

In this step the data mining tool, Weka is selected to describe the appropriate algorithms to be implemented in the study since the algorithms that are used is supported the software of choice. Weka software settings allow us to select algorithms for a mining task. Many data-mining algorithms are available for a given task. Decision tree J48 algorithm is used in this research work to show the drawbacks. Algorithms differ not only in the accuracy of their end-product, but also in the computational resources they require.

A Weka preprocessing tool is used to convert raw data into a format understandable by the data-mining algorithm.

1.5.2.5. Build data-mining model

The other important step in data mining process is building the model. The output of executing a data-mining task is the data-mining model: That model, ideally, is a representation of the data suited to our objective. One may need to explore alternative models to find the one that is most appropriate in solving the business problem. For instance, the model might be a neural network, a decision tree, or even a set of rules understandable by humans. This model building involves generating samples for training and testing the model with large data set. And finally select the best and useful model. This leads to the next step in DM process.

1.5.2.6. Test and refine the models

In order to evaluate the performance of a model before deployment, there is a need to examine the error rate on the data set that did not involve in the process of model formulation. The classifier and the models can be evaluated using evaluation criteria. The evaluation is important for understanding the quality of the model, for refining parameters in the JDM iterative process and for selecting the most acceptable model from a given set of models. There are several criteria for evaluating models. Naturally, classification models with high accuracy are considered better. This measures the true and false positive rate, the precision, and the recall of the models developed. Having high precision and recall is considered as a best model. However, there are other criteria that can be important as well.

For this particular research work the analysis of error rate generated by the confusion matrix is one mechanism used for testing the model performance. So, the key objective in this step is to determine if there is some important business issue that has been insufficiently undertaken.

The other way to evaluate models is to apply the newly gained insight to past data, and compare that with results that would be obtained without the aid of that insight, for instance by tenfold cross validation random sampling. Ideally, we newly gained insight should produce improved results—a "lift," in data-mining jargon.

Once the testing data is classified with reasonable accuracy, the rules that are required for classification can be extracted.

1.5.3. **Ethical considerations**

The study is conducted after getting approval from the Research Ethics Committee (REC) of the College of Health Sciences, Addis Ababa University. As it is a secondary data, there is no contact to individuals whose records analyzed. However, the confidentiality of data is maintained. That is, there is no data transfer to third party and using the data for other purpose beyond to this research work.

1.6. **Significance of the Research**

The need to analyze vital statistics database system of BRHP, aim to determine the internal/external requirements of the database, data dictionary and other form to determine the information contents for database and data-processing requirements for users.

This paper aims to highlight the important role of DM in analyzing the vital information, and propose a basic process of vital information analysis based on DM, in order to explore the new ideas of excavating useful intelligence based on data mining methods: Firstly, collect the vital information related to the analyzing technology field from the BRHP database, and then utilize data mining technique to extract useful new knowledge by applying different methods and algorithm to mine and analyze the data deeply.

There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just concerns of public health but also the private health sector.

- There is a wealth of knowledge to be gained from computerized health records.
- There is evidence-based medicine and prevention of hospital errors.
- When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases.
- By mining health records, such safety issues could be flagged and addressed by health management and government regulators.
- Using data mining, public health practitioner is able to discover patterns among health centers that lead to policy recommendations to their Institute of Public Health.
- They might conclude that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making."

By using data mining and visualization, public health experts could find patterns and anomalies better than just looking at a set of tabulated data.

From these necessitated ideas and issues, the discovery of new knowledge from vital statistics database (VSD) has heralded a new era in vital event registration practice nationwide with emphasis on two fundamental objectives, quality and prevention of population health problems. Health quality remains an issue of major concern in public health in Ethiopia where national vital event registration services and policies, appropriate infrastructure, trained personnel and financial resources are inadequate.

1.7. **Organization of the Paper**

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

The second chapter is dealt about literature review on data mining technology, methods/techniques used, and its application in the health care sector.

The third chapter is devoted to give understanding about the data mining tool, techniques and algorithms that applied in the study. In this chapter, issues related to data analysis tool, classification technique, and how the algorithm under the technique work for the BRHP database is addressed.

The fourth chapter is discussed about the selection and preparation of data process that is undertaken in the research work. This chapter is mainly used for understanding the process in preparing the data for producing quality data using the Weka filter options and the statistical software like SPSS tool.

The fifth chapter is provided discussions about the experimentation and result analysis in different data mining steps that is undertaken in the research work. This chapter is mainly used for describing the experiment in the data prepared till discussing results. This includes training, model building and testing results obtained by using Weka software. Results are also analyzed and interpreted.

Finally, chapter six provides concluding remarks and pointers for future work to explore and discover knowledge on health care data in general and vital statistic data in particular.

CHAPTER TWO

LITERATURE REVIEW

With the evolution of machines, we have found that some tiring and routine or complex mathematical calculations can be done using calculators, finding specific information in a large database can be done using machines fast and easily. We use machines for storing information, remind us of appointments, and so on. As the size of the data was increasing computer storage has increased. Due to the vast amount of data that was being created humans invented algorithms that produce results once a query is supplied. Although these tools perform very well, they can be used to perform only routine tasks. Automatic classifications cannot be done using standard database languages. This has led to the creation of machine intelligence algorithms that perform tasks supplied by humans and make decisions with little human supervision [3].

From the evolution of machine intelligence came data mining. In data mining, algorithms seek out patterns and rules within the data from which sets of rules are derived. Algorithms can automatically classify the data based on similarities (rules and patterns) obtained between the training on the testing data set [18].

Today, data mining has grown so vast that they can be used in many applications; examples include predicting of corporate costs claims, risk management, financial analysis, insurance, process control in manufacturing, in healthcare, and in other fields [19].

2.1. Data Mining and Its Importance

Data mining is the science and technology of exploring data in order to discover previously unknown patterns. It can be defined as the process of the non-trivial extraction of implicit, unknown, and potentially useful information from data [3]. Data mining as automated pattern recognition is a set of methods applied to knowledge discovery that

attempts to uncover patterns that are difficult to detect with traditional statistical methods. Patterns are evaluated for how well they hold on unseen cases.

Mining of data in general terms can be elaborated as retrieving useful information or knowledge for further process of analyzing from various perspectives and summarizing in valuable information to be used for increasing revenue, cut cost, to gather competitive information on business or product [20]. Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data [21].

Nowadays, the world is regarded as an expanding universe of data. We have rather much information, but limited knowledge. Some people look at this phenomenon as a new paradox of the growth of data, that is, more data means less information. Therefore, there is an urgent need for the development of new techniques to find the required information from huge amount of data [22].

With the help of data mining methods, useful patterns of information can be found within the data, which can be utilized for decision making or problem solving . Data mining is often overlooked when in fact it can provide very interesting information that statistical methods are unable to produce or produce properly. These data mining methods give us a lot more control.

The other question that arises is how to classify this massive amount of data. Automatic classification is done based on similarities present in the data. The automatic classification technique is only proven fruitful if the conclusion that is drawn by the automatic classifier is acceptable to the end user. The data we have is often vast, and noisy, meaning that it's imprecise and the data structure is complex. This is where a purely statistical technique would not succeed, so data mining is a solution.

Accurate data mining solutions could prove to be an effective way to cut down cost by concentrating on right place [20]. There are various factors that make data mining as a very important technique. First, data mining algorithms can find "optimal" interesting regularities in a database. Second, data mining algorithms typically zoom in on

interesting sub-parts of the databases. Thus, with the help of machine learning techniques data mining make it easier to find interesting connections in Database.

In short, with a computer, data mining can guide you to automatically find the one "information diamond" among the tons of data debris in the database. It seems like a very attractive research area. Especially, it shows its importance in the following business area [22].

1. Market Management: This includes target marketing, customer relationship management, market basket analysis, cross selling, market segmentation.
2. Risk Management: This includes forecasting, customer retention, improved underwriting, quality control, competitive analysis.
3. Fraud Management: The concern of fraud management is fraud detection and prediction. That is, identify loyal' one in like customers service provisions.
4. Text Mining: The application of traditional data mining technology to the unstructured contexts of text Databases.
5. Web Analysis: Use data mining technology in conjunction with the Internet.

Since data mining has its unique importance in a large area, it is very important in the whole process of KDD. The accessibility and abundance of information today makes data mining a matter of considerable importance and necessity. Together with the highly importance of this new technology, efforts have been made to define standards for data mining process. For example, CRoss Industry Standard Process (CRISP-DM 1.0) in 1999 and Java Data Mining standard (JDM 1.0) in 2004 have been developed to avoid lack of uniformity in the data mining process. Independent of these standardization efforts, freely available open-source software systems like the RProject, Weka, KNIME, RapidMiner, jHepWork and others have become an informal standard for defining data-mining processes [7].

In general the following life cycles have been existed as they represent the most authoritative, most cited, and most applied life cycles in both academia and industry:

- Knowledge Discovery in Databases Process [4, 23]

- Refined KDD Paradigm [24]
- Knowledge Discovery Life Cycle (KDLC) Model [25]
- Information Flow in a Data Mining Life Cycle [26, 27]
- CRoss-Industry-Standard Process for Data Mining (CRISPDM) [28]
- Java Data Mining API (JDM) [7]

Some of the life cycles only differ marginally whereas others differ considerably in structure and comprehensiveness. All, but the CRISP-DM life cycle have been developed by academia. Although data mining has been a very evolving research area in recent years, there is still a need for a comprehensive and complete life cycle. The CRISP-DM reference model is the most complete and also the most widely accepted and applied data mining life cycle [29].

As we can observe from the following figure 2.1, a typical data-mining system consists of a data-mining engine and a repository that persist the data-mining artifacts, such as the models, created in the process. Due to its latest in technology compared to the CRISP-DM the researcher interested to use the JDM model in order to achieve the objectives of the study. A key JDM API benefit is that it abstracts out the physical components, tasks, and even algorithms, of a data-mining system into Java classes [30].

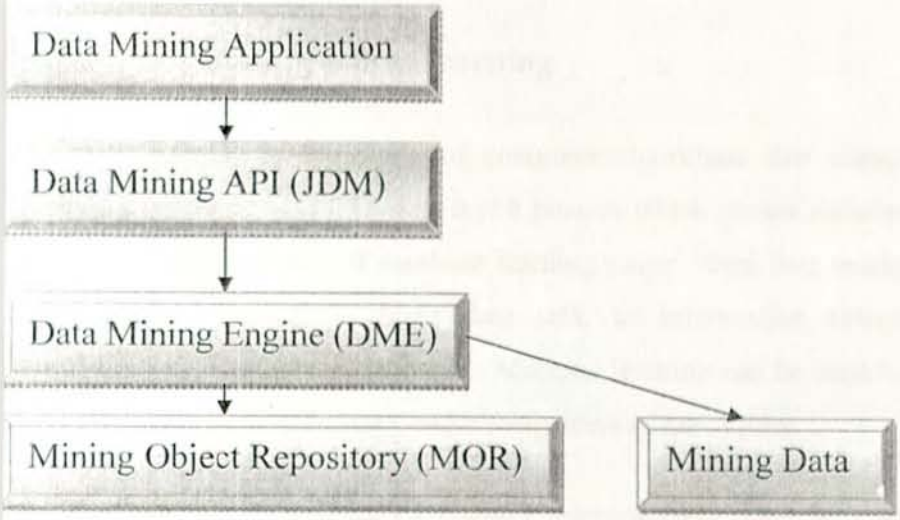


Figure 2.1. Components of a data-mining system.

Building a data-mining model typically starts with identifying recurring patterns in the data, and then distilling those patterns in a way that helps communicate them to humans or other machines. Models can take the form of a graphical representation, a set of equations, a neural network, or even a collection of rules. Models can be applied to new data, or evaluated and refined in the presence of ever larger data sets.

The current trend is towards automating as much of this process as possible such that even those not expert in data mining can reap the benefits of data-mining technologies[30].

2.2. Multidisciplinary Nature of Data Mining

Data mining has its origins in conventional artificial intelligence, machine learning, statistics, and database technologies, so it has much of its terminology and concepts derived from these technologies [31]. As data mining is an interdisciplinary field [32], it incorporates many different approaches, technologies, and methodologies to be able to generate and discover new and innovative knowledge. This interdisciplinary approach involves tools and models from statistics, artificial intelligence, pattern recognition, heuristics, data acquisition, data visualization, optimization, information retrieval, high end computing, and others [21, 33].

2.2.1. Machine learning

Machine Learning is the study of computer algorithms that improve automatically through experience [18] . That is, it is a process which causes systems to improve with experience. Applications of machine learning range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. Machine learning can be used to develop systems resulting in increased efficiency and effectiveness of the system.

Machine learning is also called concept learning [4]. That is, computers can learn concepts and patterns within the data. Concept learning acquires the definition of a general category given a sample of positive and negative training examples of the

category, the method of which is the problem of searching through a hypothesis space for a hypothesis that best fits a given set of training examples. Machine learning is considered successful when it can correctly find all the instances that consist of the right patterns and concepts, although at times a machine cannot categorize correctly all the instances due to high variations in attributes present in the data [9].

The two important areas of application in machine intelligence are knowledge discovery and classification [34]. Knowledge discovery is defined as the non-trivial extraction of implicit, unknown, and potentially useful information from data. Classification is probably the oldest and most widely-used of all the KDD approaches [35]. Classification is learning a function that maps (classifies) a data item into one of several predefined classes [19]. Patterns that are extracted using machine intelligence can be used to predict which class the data falls under

A decision support system is similar to a machine learning system [36]; it is a system that suggests decisions based on the patterns found in the data. The three components required for decision support systems are the end user, hardware and software products, and data mining process that interpret and discover knowledge necessary for decision making.

2.2.2. Statistics and Data Mining

The two disciplines 'Statistics' and 'Data mining' are very similar. Statisticians and data miners commonly use many of the same techniques. Statistics developed as a discipline separate from Mathematics over the past century and a half to help scientists for making some sense of observations and to design experiments that yield the reproducible and accurate results associated with the scientific method. For almost all of this period, the issue was not too much data, but too little. Whereas, Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Or, data mining is the application of Statistics in the form of exploratory data analysis and prediction models to reveal patterns and trends in very large datasets [15].

Traditionally, analysis was strictly a manual process. The traditional method of turning data into knowledge relies on manual analysis and interpretation [37]. In normal statistics, one is trying to relate two things (say, smoking and cancer rates) and prove whether or not that link is true. One or more analysts would become intimately familiar with the data and -- with the help of statistical techniques -- provide summaries and generate reports. Statistics has a solid theoretical foundation but the results from statistics can be overwhelming and difficult to interpret as they require user guidance as to where and how to analyze the data. Statistical analysis systems such as SAS and SPSS have been used by analysts to detect unusual patterns and explain patterns using statistical models such as linear models. Such systems have their place and will continue to be used [15, 22].

In fact, such manual data analysis is becoming impractical in many domains as data volumes grow exponentially. Databases are increasing in size in two ways: the number of records, or objects, in the database, and the number of fields, or attributes, per object [38].

Such an approach rapidly breaks down as the quantity of data grows and the number of dimensions increases, especially, when there are millions of cases and each having hundreds of fields. In this situation, we need the data mining to help us automate data analysis process to derive some useful hidden information. Data mining allows the expert's knowledge of the data and the advanced analysis techniques of the computer to work together. The time to use data mining technique really depends on the user's need. Basically, whenever the users want to extract information useful for decision support or exploration and understanding the phenomena governing the data source, they urgently need the help of the data mining [37].

Data mining is about using statistical and machine learning techniques to find relationships in a set of large and complex data [22].

In data mining, a researcher attempt to find interesting relationships that we may not even know exist – we are just hoping that if one look at enough information, certain links (e.g.

that people who smoke a lot are more likely to get cancer) may pop out of it. Once a possible relationship is found, we'll test them on other data to see if they're true or not.

Data mining is normally used when one don't know what we are looking for. Stock market analysts like to use data mining to try and find patterns - such as stock prices falling after floods.

Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large Databases. Basically, it is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It refers to the application of algorithms for extracting patterns from data. It is an important stage in knowledge discovery process. In general, knowledge discovery process includes six stages such as data selection, cleaning, enrichment, coding, data mining, and reporting. At each stage, the data miner can step back one or more phases. In this sense, data mining refers to a class of methods that are used in some of the steps comprising the overall Knowledge Discovery in Database (KDD) process.

Many of the data mining techniques were invented by statisticians or have now been integrated into statistical software; they are extensions of standard statistics. Although, data miners and statisticians use similar techniques to solve similar problems, but the data mining approach differs from the standard statistical approach in several areas such as [39]:

- Data miners assume that there is more than enough data and processing power.
- Data mining assumes dependency on time everywhere.
- It can be hard to design experiments in the business world without data mining.

Because of the nature of the problems, there are some differences (rather than opposites) in approaches that the statisticians and the data miners used to solve similar problems. As such, they shed some light on how the business problems addressed by data miners differ from the scientific problems that spurred the development of statistics. One major

difference between business data and scientific data is that the latter is non-truncated /non-censored data and the former is truncated /censored data. Given a methodology or an algorithm to analyze data, it is often very hard to say whether it is "Statistics" or "Data Mining". It is not clear how one should put this label. Actually, in practice, while dealing with the real life problems in the industry, customers never ask, "Are you a data miner or a statistician?" In fact their main interest is in solving the problem in hand to their level of satisfaction and it is immaterial what label they use. As service providers to the customers, a data miner or a statistician need to try those statistical techniques or algorithms in the bag, which best suited to answer the queries of the customer [15].

As in [22, 38] described, figure 2.2 and figure 2.3 summarized the stages/processes identified in data mining and statistics to show the difference between them. In statistics, the process works as follows:

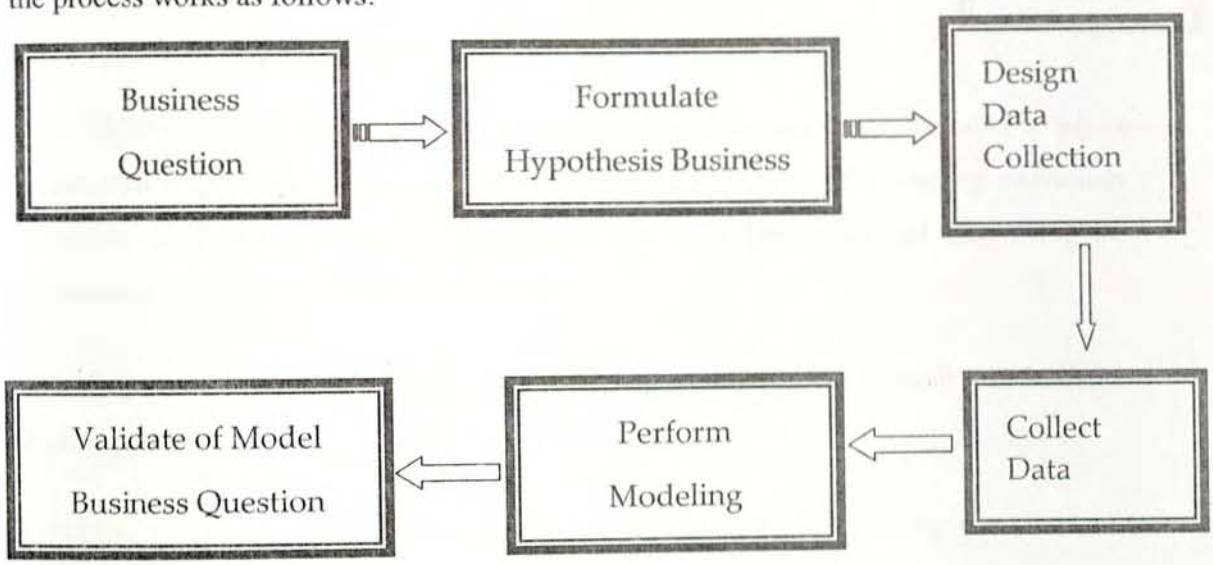


Figure 2.2 Statistical analysis steps

As most of the problems can be downsized to classical statistical techniques, the DM expert should have a sound knowledge of statistics (and the corresponding SAS procedures) and how to use them. Data mining claims a novel kind of data exploitation - it is not simply the hypothesis confirmation of statistics; nor is it simply the data visualization of graphs and plots. Data mining is becoming a force to be reckoned with, because of the way it can generate new ideas.

Hence, data mining is carried out as:

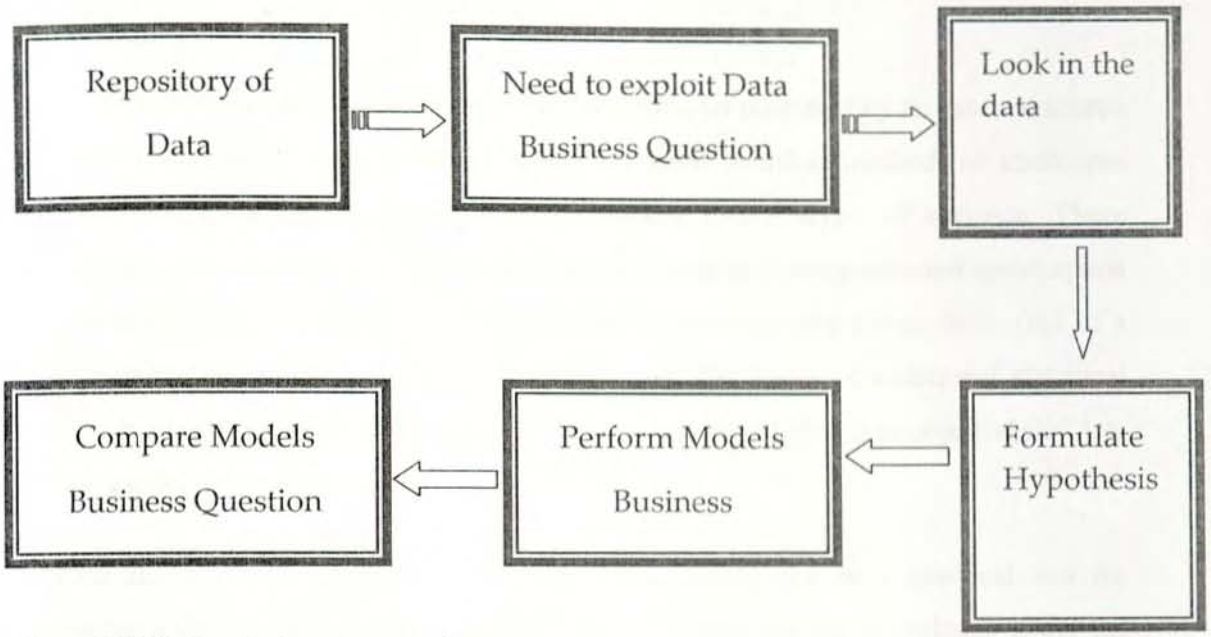


Figure 2.3 Data mining process steps

Like Statistics, Data Mining is not only modeling and prediction but also a whole problem solving process. In short, data mining is the process of extracting previously unknown, valid and actionable information from large Databases and then using the information to make crucial business decision.

Following are some examples of (possible) situations where Data Mining and Statistics can be used.

Normally, for proving the efficiency of a drug, the rules for playing the game are described. Take for example a drug for hay fever, where one knows how to measure (relief) to compare the drugs. With Data Mining techniques, we could try to find alternative measures of relief, and promote the drug in another way: on top of curing the disease in a standard way, with drug you get some extras compared to the competitor, e.g. a better Quality of Life

According to Last and Kandel [33] 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.

In such statistical methods the objective of a statistical test is to verify the null hypothesis.

Verification methods deal with evaluation of a hypothesis proposed by an external source (like an expert etc.). These methods include the most common methods of traditional statistics, like goodness-of-fit test, t-test of means, and analysis of variance. These methods are less associated with data mining than their discovery-oriented counterparts because most data mining problems are concerned with selecting a hypothesis (out of a set of hypotheses) rather than testing a known one. The focus of traditional statistical methods is usually on model estimation as opposed to one of the main objectives of data mining: model identification [40].

As Last and Knadel [33] 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".

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 [33]. However, it does not mean that data mining has replaced other statistical methods such as OLAP, Regression, etc. Rea [41] 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".

2.3. Knowledge discovery and data mining

The traditional method is used to analyze data manually for patterns for the extraction of knowledge. Take any field like banking, healthcare, and marketing; there will always be a data analyst to work with the data and interpret the final results. For example, in the case of health care, the health organizations analyze the trends in diseases and the occurrence rates. This helps health organizations take precautions in decision making and planning of health care management. The analyst acts like an

interface between the data and knowledge. We can use machine intelligence to assist the analyst to produce similar results or knowledge from the data [42].

Discovering patterns within a database is usually called data mining or knowledge extraction. The term data mining is used mostly by statisticians, data analysts and the management information systems (MIS) [43].

Concerning the two terms knowledge discovery and data mining, there are different conceptual views that some author [15] differentiate knowledge discovery in database (KDD) as the whole process and while others [44] views as they are one and the same. Some of the numerous definitions of Data Mining, or Knowledge Discovery in Databases are: Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

As defined by Fayyad et al [4] "Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data Mining (DM), on the other hand, defined as the application of algorithms for extracting patterns from data without the additional steps of the KDD process. Other definitions for data mining are given by Berry and Linoff [15].

The analogy with the mining process is described as Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. 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. While Data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

From these ideal perspectives, we reveal as KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also

includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

The difference between knowledge discovery and data mining is that the latter is the application of different intelligent algorithms to extract patterns from the data whereas knowledge discovery is the overall process that is involved in discovering knowledge from data. There are other steps such as data preprocessing, data selection, data cleaning, and data visualization, which are also a part of the KDD process [19].

Data Mining can be considered as a central step of the overall process of the Knowledge Discovery in Databases (KDD) process. Due to the centrality of data mining in the KDD process, there are some researchers and practitioners that use the term "data mining" as synonymous to the complete KDD process [40].

As stated by the Java data mining (JDM) standard process [31], the overall process of finding and interpreting patterns from data involves the following steps for the data mining process.

1. **Select data mining objective:** - The first, and most important, step is to decide what kinds of new knowledge or insight we want to gain from the data. The more specific we are, the more likely our data-mining process succeeds.

   In this regard, the researcher makes efforts by reading different published articles and consulting the concerned people to understand the business settings and to gain knowledge for the problem under study as stated in detail at this chapter. In line with this, a data mining research problem is planned with set of objectives to approach the problem under the first chapter of the paper.

2. **Select and prepare data:** - As per JDM the second step involve data understanding. Once we've decided on the objective, we must identify the data that we think may help we achieve those goals. Initially, we may wish to select only the subset of the available data that we believe is most representative of what

we wish to find out. We can later select additional data subsets to improve our initial findings.

Relevant data sets are seldom in the format suited to data mining. Often, we must transform that data, possibly cleaning it—eliminating incomplete records, for instance—and sometimes also pre-processing it. This endeavour leads to initiate data preparation. This is often the most time consuming task of DM processes, especially if data is drawn directly from the company's operational database rather than data warehouse.

So, the primary data source for this study is the secondary data accumulated in the Butajira Rural Health Program (BRHP) database, which contains epidemiology data. Hence to detect the data quality problems and to identify interesting subsets of the data different literatures and previous studies have been explored. In general this phase covers all activities to make ready the final data set from the initial row data for model building and analysis purpose. A pre-processing tool is used to convert raw data into a format understandable by the data-mining algorithm. Details of these activities are presented in chapter four.

3. **Choose and configure the mining tasks.** Next, we should decide on the specific data-mining task to perform. For instance, we may wish to cluster households together that are visited similar cases, and then derive classification rules that predict how those mortality is classified. Those rules, in turn, can help evaluate and decide what new for future on those households.

   Having selected a mining task, we would then configure that task with parameters suitable for the task. In the JDM API, such configuration is specified with *settings*. There are a variety of DM tasks available, all with pros and cons, depending on the business problem at hand and the data available for analysis. Most are based on statistical or computer science algorithms. Some of the more common techniques include classification, clustering, and association rules. Thus, we concentrate mainly on supervised learning methods in data mining.

4. **Select and configure the mining algorithms.** In this step the data mining tool, Weka is selected to describe the appropriate algorithms to be implemented in the study since the algorithms that are used is supported the software of choice. Weka software settings allow us to select algorithms for a mining task. Many data-mining algorithms are available for a given task. Algorithms differ not only in the accuracy of their end-product, but also in the computational resources they require.

   Many data-mining tools are able to automatically match algorithms to a desired data-mining objective; for instance, a clustering algorithm to create data clusters, or an association-rules algorithm to identify association rules. Under this, we look at a data-mining algorithm (decision tree classification) to perform automatic classification based on the testing data set and also provide accuracy in terms of percentage with regard to the number of cases in the testing dataset, which are classified correctly.

5. **Build data-mining model.** The other important step in knowledge discovery process is building the model. The output of executing a data-mining task is the data-mining model: That model, ideally, is a representation of the data suited to our objective. One may need to explore alternative models to find the one that is most appropriate in solving the business problem. For instance, the model might be a neural network, a decision tree, or even a set of rules understandable by humans. This model building involves generating samples for training and testing the model with large data set. And finally select the best and useful model. This leads to the next step in DM process.
6. **Test and refine the models**. Evaluating the performance of a model is a fundamental aspect of machine learning. For example, a classifier receives a training set as input and constructs a classification model that can classify an unseen instance. Based on this, several models might be created and needed to evaluate the accuracy of each model with past data, and possibly to select a "best" model for the purpose of the study setup.

In order to evaluate the performance of a model before deployment, there is a need to examine the error rate on the data set that did not involve in the process of model formulation. The classifier and the models can be evaluated using evaluation criteria. The evaluation is important for understanding the quality of the model, for refining parameters in the KDD iterative process and for selecting the most acceptable model from a given set of models. There are several criteria for evaluating models. Naturally, classification models with high accuracy are considered better. This measures the true and false positive rate, the precision, and the recall of the models developed. Having high precision and recall is considered as a best model. However, there are other criteria that can be important as well.

For this particular researcher work the analysis of error rate generated by the confusion matrix is one mechanism used for testing the model performance. So, the key objective in this step is to determine if there is some important business issue that has been insufficiently undertaken.

The other way to evaluate models is to apply the newly gained insight to past data, and compare that with results that would be obtained without the aid of that insight, for instance by tenfold cross validation random sampling. Ideally, we newly gained insight should produce improved results—a "lift," in data-mining jargon.

Lastly, after the complete DM process stages and final result obtained the researcher can present to the department since the study is conducted for academic achievement and report the findings to subject matters expert consultation for validity and acceptance in the study domain. In some cases, we may build systems that automatically improve their data-mining models with new data, or systems that take actions in the presence of a continuous stream of new information.

2.4 Data Mining Tasks

The data mining stage of KDD process consists of verification oriented (the system verifies user's hypothesis) and discovery oriented (the system finds new rules and

patterns autonomously) [4]. According to Fayyad et al [4] the discovery methods are methods that automatically identify patterns in the data. The discovery-oriented methods can be further partitioned into descriptive (e.g. Clustering, Summarization, Visualization) and predictive (like classification, regression, etc). Description-oriented Data Mining methods focus on (the part of) understanding the way the underlying data operates, where prediction-oriented methods aim to build a behavioural model that can get newly and unseen samples and is able to predict values of one or more variables related to the sample. However, some prediction-oriented methods can also help provide understanding of the data [40].

Fayyad et al [4] note that these are two "high-level" primary goals of data mining practice: Prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Often the emphasis of predictive modelling is on predictive accuracy rather giving more emphasis on understanding the model. However, description focuses on finding human-interpretable patterns describing the data.

Most of the discovery-oriented techniques are based on inductive learning [40], where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future unseen examples. Strictly speaking, any form of inference in which the conclusions are not deductively implied by the premises can be thought of as induction.

As we said before data mining is one among the most important steps in the knowledge discovery process. It can be considered the heart of the KDD process. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the data.

The goals of prediction and description are achieved by using the following primary **data mining tasks** [4, 9, and 45]: Classification, clustering and association rule discovery.

2.4.1 Classification

Classification is categorized as supervised learning which is used to predict a value. They require a user to specify a set of predictor attributes and a target attribute. Predictors are the attributes used to predict the target attribute value. Classification is learning a function that maps (classifies) a data item into one of several predefined classes. The learning algorithms take a set of classified examples (training set) and use it for training the algorithms. With the trained algorithms, classification of the test data takes place based on the patterns and rules extracted from the training set. Classification can also be termed as predicting a distinct class. It predicts categorical class labels (discrete or nominal) [34].

Applications can select an algorithm that works best for solving a business problem. Selecting the best algorithm and its settings values requires some knowledge of how each algorithm works and experimentation with different algorithms and settings. The popular classification techniques are decision tree, neural networks and Bayesian network [45].

Decision tree is a widely used learning method [9]. Rules from the training dataset are first extracted to form the decision tree which is then used for classification of the testing dataset. A decision tree is necessarily a tree with an arbitrary degree that classifies instances. They are a powerful tool for classification and predication but require extensive computation. Creating the tree based on the training set takes time although making decisions once the tree is made is not time consuming. Classification tree algorithms may be divided into two groups: one whose result is a binary tree and other that yields non-binary trees (also called multiway) splits [12]

Neural network is another classification technique. It is often defined as a computer application that mimics the neurophysiology of the human brain; a neural network is capable of learning from examples to find patterns in data. A neural network (NN) learns by using a training set to regress through the examples and learn in a non-linear manner. The neural network is first trained, which involves reading sample data and iteratively adjusting the network's weights to produce optimum predictions [46].

Then new data can be applied to this model to quickly generate predictions. Neural networks are reputed to produce highly accurate results and, in practical applications, can contribute to profitable decisions.

The standard validation-set concept can be used to avoid overfitting of the training set. Similar to the method of moving 10% of the training examples into a tuning set for ID3 pruning, 10% of the NN training data will be moved into a tuning set to validate our learned function. After every five epochs of training we save the network weights at that time step and calculate the error on the validation set with these weights. If at any time the error rate is lower than the previous error rate from five epochs previous, training is stopped and the network weights of the previous validation step are used [46].

Neural networks and tree-based models are both effective data mining tools; however they each have their own unique strengths. Both have advantages over linear models by being able to detect nonlinear relationships automatically. But they have different virtues when it comes to making predictions from a large number of predictor variables. Tree-based models are good at selecting important variables, and therefore work well when many of the predictors are irrelevant. Neural networks are good at combining information from many predictors without over-fitting, and therefore work well when many of the predictors are partially redundant.

The Bayesian network Classifier is another type of learner [47]. It uses Baye.s theorem and assumes independence of feature values to estimate posterior probabilities.

The naïve bayes method is based on probabilistic knowledge. This method goes by the name Naïve Bayes, because it's based on Bayes's rule and "naively" assumes independence- it is only valid to multiply probabilities when the events are independent [35]. Thus the naïve bayes rule outputs probabilities for the predicted class of each member of the set of test instance. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data.

The Naïve Bayes classification reads a set of examples from the training set and uses the Bayes theorem to estimate the probabilities of all classifications. For each instance, the classification with the highest probability is chosen as the prediction class. The naïve Bayesian classifier traditionally makes the assumption that a single Gaussian distribution generates numeric attributes [48]. Two types of Naïve Bayes algorithms are mentioned: naïve Bayes (NB) and simple Naïve Bayes (SNB).

The difference between the two is that in NB the probability of the attributes are calculated based on normal distribution's mean, standard deviation, weighted sum, and precision but SNB is only based on mean and standard deviation.

2.4.2 Clustering

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 exclusive set of clusters. Clustering according to similarity is a very powerful technique; the key to it is being to translate some intuitive measure of similarity into quantitative measures [22].

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. Clustering come under unsupervised (undirected) category. That is, the system clusters the data into their natural group/category. The system has to discover subsets of related objects in the training set and then it has to find descriptions that describe each of these subsets [4].

Unsupervised functions are used to find the intrinsic structure, relations, or affinities in data. Unsupervised mining doesn't use a target field. Clustering is used to find the natural groupings of data. That is, the grouping of similar instances in to clusters takes place. The challenges in this type of machine learning technique is that we have to first identify clusters and assign new instances to these clusters. The relationship among the data is identified by bottom-up approach (which is discussed after section 2.4.3) [9, 45].

There are a number of approaches for forming clusters. One approach is to form rules which dictate membership in the same group based on the level of similarity between

members. Another approach is to build set functions that measure some property of partitions as functions of some parameter of the partition. Two methods are often used in database clustering. This includes Partitioned algorithms and hierarchical algorithms [22].

Cluster analysis is the problem of decomposing or partitioning a (usually multivariate) data set into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups [38]. Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors. The two typical methods are distance-based: K-means clustering and model-based: expectation maximization (EM) clustering.

K-means clustering is an iterative algorithm, which starts centroid with random cluster centers. A single iteration assigns all objects to the closest clusters based on their distances from the cluster means and then recomputes the cluster means.

K-means clustering [7, 49] which is a simple technique to group items into k clusters. It works with numeric data only.

If we need to group categorical data we employ hierarchical clustering technique. It is a second important category of clustering method [16]. Hierarchical clustering algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

The most common distinction between the two types of clustering techniques is whether the set of clusters are nested or unested. That is, a partional clustering is simply grouping set of data objects into non-overlapping subsets such that each data objects is in exactly one subset, while hierarchical clustering is a set of nested clusters that is organized as a tree. Each node in the tree is the unions of its children and the root of the tree is the cluster all the objects [16].

2.4.3 Association Rule

Similarly, association functions also come under the unsupervised category. It is used to infer co-occurrence rules from the data. Association rule is also known as market basket analysis. It discovers interesting associations between attributes contained in a database. In pattern discovery, two steps can be took place [22]. That is, fining frequent patterns from large item sets and generating association rules from these item sets. Frequent pattern is a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set. Naïve algorithm, Apriori and FPGrowth are the frequent item set mining methods.

Most algorithms such as Apriori for the discovery of large item set work as follows [22]. First, the supports for single items are computed and large 1-itemsets are found. Then, iteratively for sizes s=2, 3..., candidate s-itemsets are generated from the large (s-1)-itemsets of the previous pass. Supports for the candidates are then computed from the database, and those candidates that are turned out to be large are used in next pass to generate candidates of size s+1.

Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X==>Y, where X and Y are sets of attributes, meaning that in the rows of the database where the attributes in X have value true, also the attributes in Y tend to have value true. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y.

According to Han and Kamber [21] in order to determine the best technique suitable for specific data mining problem, the two styles of data mining are directed and undirected. Directed Data Mining is a top-down approach. It is used to predict when we know approximately what we are looking for or what we want As it is a predictive model, it uses experience to rank possible outcomes in the future by calculating a score for each outcome. The model is seen as a black box because we care only about the predictions and not how it actually works. Building a predictive model is to apply knowledge gained

in the past to the future. Which customers are likely to buy a specific type of car or else? This could be considered as an example.

Meanwhile, undirected data mining follows bottom-up approach. It finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important. Under this approach, we want to know how the model works and how it comes up with the answer. Human interaction is necessary because only people can determine what significance, if any, the patterns have. It is often used during the data exploration steps. For example, a person looks at a decision tree and possibly notices an interesting pattern. However, decision tree could make predictions for directed data mining.

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 [50]. The quality of the models/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 models/rules.

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.

2.5 Data Mining for Health Informatics

Healthcare is a very research intensive field and the largest consumer of public funds. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no longer ignore these emerging tools. This resulted in uniting of healthcare and computing to form health informatics (Health informatics exists since the 1950's). Health informatics plays a very important role in the use of clinical data. This is expected to create more efficiency and effectiveness in the health care system, while at the same time, improve the quality of health care and lower cost[51].

Health informatics is an emerging field. It is the logic of healthcare. It is the field that concerns itself with the cognitive, information processing, and communication tasks of medical practice, education, and research, including the information science and the technology to support these tasks [42].

Health informatics is especially important as it deals with collection, organization, storage of health related data. With the growing number of patient and health care requirements, having an automated system will be better in organizing, retrieving and classifying of medical data. Physicians can input the patient data through electronic health forms and can run a decision support system on the data input to have an opinion about the Parent's health and the care required. An example in the advances in health informatics can be the diagnosis of a patient is health by a doctor practicing in another part of the world. Thus healthcare organizations can share information regarding a patient which will cut costs for communication and at the same time be more efficient in providing care to the patient [9].

There are other issues like data security and privacy, which is equally important when considering health related data. Thus Health informatics "deals with biomedical information, data, and knowledge--their storage, retrieval, and optimal use for problem solving and decision making"[42]. This is a highly interdisciplinary subject where fields in medicine, engineering, statistics, computer science and many more come together to form a single field.

Health Informatics comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and health care.

With the help of smart algorithms and machine intelligence, one can provide quality of healthcare, problem solving skills and decision-making systems. Information systems can help in supporting clinical care in addition to helping administrative tasks. Thus the physicians will have more time to spend with the patients rather than filling up manual forms.

First the paper forms that are filled by the physicians are converted into electronic forms. Programs can be built around these forms to help in input validations. Some of the validation steps can be in the form of cautions provided when fields are inputted with invalid values; another type of validation can be to make sure attributes of high priority are not left empty by the user.

The informatics part of health care can take care of the structuring; searching, organizing and decision making with the emergence in health informatics came many important research ideas and fields of study. The discipline utilizes the methods and technologies of the information, social and technology sciences for the purposes of problem solving and decision-making thus assuring quality healthcare in all basic and applied areas of medical, biomedical and health sciences.

Health informatics is concerned primarily with the processing and dissemination of data, information and knowledge in all aspects of healthcare. It aims to study the principles and provide solutions. Domains of health informatics include research, academia, operations and commercial.

As a discipline it is used by clinicians, operational health practitioners, managers, academics, researchers, educators, scientists, technologists, and political leaders [52]. Regarding these technologies, the good news is that they have become relatively inexpensive (e.g., hundreds of dollars per sample), making them widely accessible to researchers. However, the bad news to many medical researchers are that the amount of data collected by these techniques is phenomenal.

The ability to use data to extract useful information for quality healthcare is crucial. Health informatics plays a very important role in the use of clinical data for discoveries of pattern that is important for the diagnosis of new diseases. Computer assisted information retrieval may help support quality decision making and to avoid human error. This need lead to the use of data mining in Health informatics. Data mining techniques, hence, become attractive for many medical and health studies. Broadly

defined, there are four general objectives for the data mining activities in health informatics [53].

The first one is diagnostics to determine whether a patient is suffering from a certain medical condition. For instance, early stage lung and oral cancers are very hard to diagnose by conventional means; genomic signatures can be used to provide more timely and perhaps more accurate diagnosis. The second is prognostics to predict how well a patient would re- cover or how the medical condition would progress over time. For instance, biomarkers have been identified to predict how well a transplanted organ would be tolerated in the recipient's body.

The third one is treatment optimization to predict the response to treatment or therapies. For example, for certain cancer types, biomarkers can be used to predict whether a certain chemotherapy regimen would be effective or not. Pharmaco-genomics is a very active area of research on understanding how pharmaceuticals and medications can affect the genomic profile of a patient. Understanding of disease mechanisms is the fourth to provide new insights into how a certain medical condition is triggered. For example, it is an active area of research on finding out how signaling pathways interact during a viral infection.

2.6 Data mining Practices in Health care

The practice of using concrete data and evidence to support medical decisions (also known as evidence-based medicine or EBM) has existed for centuries. John Snow [54], considered to be the father of modern epidemiology, used maps with early forms of bar graphs in 1854 to discover the source of cholera and prove that it was transmitted through the water supply.

Snow counted the number of deaths and plotted the victim's addresses on the map as black Bars to discover that most of the deaths clustered towards a specific water pump.

Similarly, Florence Nightingale invented polar-area diagrams in 1855 to show that many army deaths could be traced to unsanitary clinical practices and were therefore

preventable. She used the diagrams to convince policy-makers to implement reforms that eventually reduced the number of deaths [55].

Due to such need of knowledge from data, real strategic value comes from understanding customer behavior and being able to model alternative actions. The knowledge required to anticipate behavior could be discovered from many users running several traditional queries against data warehouses, but that supposes the questions are known and the time is available to complete the analysis.

According to Benzler, et al. [56] explanation as we enter the new millennium, the revolution of the information age still gaining speed, it seems inconceivable that large parts of the Earth's population remain devoid of vital health information. For one billion people living in the world's poorest countries, where the burden of disease is highest, no one registers those who are born or who die or ascertains the causes of their deaths. From the limited data available, the health profile of these populations can be likened to an iceberg: the bulk of reliable data on trends in age, gender, geographic variations, and burden of disease remains hidden. This great void in population-based information constitutes a major and long-standing constraint on the articulation of effective policies and programs to improve the health of the poor and thus perpetuates profound inequities in health. The need to establish a reliable information base to support health development has never been greater.

Benzler, et al further stated that experience has recently emerged from a growing number of community-based field stations that have continuous monitoring systems for geographically defined populations. These field stations generate high-quality, population-based, longitudinal health and demographic data with the potential to fill this information void in the developing world. Since 1997 a number of organizations have made a systematic effort to harness and make more readily available the products of these disparate initiatives.

Snow and Nightingale [56] were able to personally collect, sift through and analyze the mortality data during their times when the volume of information was manageable.

Today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did.

This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector.

Data mining is a powerful new solution to information overload. It enables an organization to better understand the business process at work by searching automatically through huge amounts of data, looking for patterns of events, and presenting these to the business in an easy-to-understand graphical form. These systems are tireless, they do not forget, they free up skilled human resources, and find answers to important questions that users may never have asked.

Organizations need these new solutions if they are to remain competitive. According to Eight Trends in IT [57], in the Gartner Research Review from 1998, trend one is: "From Data to Decisions. Rather than using IT as a means to collect and present data for users to make decisions, technology will continue to automate more of the burden of the decision making process itself (e.g., through data mining, expert systems, and agents)."

Data mining and its application to medicine and public health is a relatively young field of study. In 2003, Wilson et al [49] began to scan cases where KDD and data mining techniques were applied in health databases. They found confusion in the field regarding what constituted data mining. "Some authors refer to data mining as the process of acquiring information, whereas others refer to data mining as utilization of statistical techniques within the knowledge discovery process."

Despite the differences and clashes in approaches, the health sector has more need for data mining today. There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just concerns of public health but also the private health sector.

Data overload: There is a wealth of knowledge to be gained from computerized health records. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge [58].

In fact, some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information. Computers and data mining are best-suited for this purpose [59].

Evidence-based medicine and prevention of hospital errors: When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. For instance, an ongoing study on hospitals and safety found that about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors [60]. By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

Policy-making in public health: Lavrac et al. [61] combined GIS and data mining using among others, Weka with J48 (free, open source, Java-based data mining tools), to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making."

Quality Healthcare is the need of an increasingly discerning population worldwide. DM in Healthcare provides comprehensive healthcare of highest quality in the geographies they operate, with an aspiration to be a responsible player with a quest for excellence [62].

The improvement of new technologies rise data collection and accumulation. Without appropriate processing and interpretation this information remains useless. Therefore, the

DM technologies give a much better and more exact representation of relationship between symptoms and diagnosis after exploring the accumulated data.

2.6.1 Challenges and concerns

Applying data mining in the medical field is a very challenging undertaking due to the idiosyncrasies of the medical profession. Shillabeer and Roddick's [59] work cite several inherent conflicts between the traditional methodologies of data mining approaches and medicine. In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practice, which simply starts with the data set without an apparent hypothesis.

Also, whereas traditional statistics is concerned about patterns and trends in data sets, data mining in medicine is more interested in the minority that do not conform to the patterns and trends. What heightens this difference in approach is the fact that most standard data mining is concerned mostly with describing but not explaining the patterns and trends. In contrast, medicine needs those explanations because a slight difference could change the balance between life and death.

For example, anthrax and influenza share the same symptoms of respiratory problems. Lowering the threshold signal in a data mining experiment may either raise an anthrax alarm when there is only a flu outbreak. The converse is even more fatal: a perceived flu outbreak turns out to be an anthrax epidemic [63]. It is no coincidence that we found that, in most of the data mining papers on disease and treatment, the conclusions were almost-always vague and cautious. Many would report encouraging results but recommend further study. This failure to be conclusive indicates the current lack of credibility of data mining in these particular niches of healthcare.

The confusion about the definition of data mining also complicates the issue. For example, we found a couple of papers with the keywords "data mining" in their titles but turned out to be the simple use of graphs. Shillabeer [59] said that this misunderstanding is prevalent in the relatively young existence of data mining in healthcare.

Even if data mining results are credible, convincing the health practitioners to change their habits based on evidence may be a bigger problem. In one case, it was found that doctors coming out of autopsy without washing hands and led to a high probability of deaths in the patients they treated after the autopsy. Presented with this evidence, doctors still refused to change their habits until only much later.

Shillabeer [59] also reported most doctors prefer to listen to a respected opinion leader in the medical profession, rather than to the result of data mining. Shillabeer's observation can be validated by us, since we have worked with doctors in a medical school in our capacity as an organizational management consultant.

Privacy of records and ethical use of patient information is also one big obstacle for data mining in healthcare. For data mining to be more accurate it needs a sizeable amount of real records.Healthcare records are private information and yet, using these private records may help stop deadly diseases.

Below are the challenges the data mining technique faces [59]:

- Large databases: Databases with hundreds of fields and tables, millions of records, and multi-gigabyte size are quite commonplace, and terabyte databases are beginning to appear. This poses a challenge that how to mine the information from such a huge database.
- High dimensionality: Not only is there often a very large number of records in database, but there can also be a very large number of fields (attributes, variables) so that the dimensionality of the problem is high. This creates problems in terms of increasing the size of the search space for model induction in a combinatorial explosive manner.
- Overfitting: When the algorithm searches for the best parameters for one particular model using a limited set of data, it may overfit the data, resulting in poor performance of the model on test data.
- Assessing statistics significance: A problem (related to overfiting) occurs when the system is searching over many possible models.

- Changing data and knowledge: Rapidly changing (non-stationary) data may make previously discovered patterns invalid. In addition, the variables measured in a given application database may be modified, deleted, or augmented with new measurements over time.
- Missing and noisy data: This problem is especially acute in business databases. Important attributes may be missing if the database was not designed with discovery in mind.
- Complex relationship between fields: Hierarchically structured attributes or values, relations between attributes, and more sophisticated means for representing knowledge about the contents of a database will require algorithms that can effectively utilize such information.
- Understand ability of patterns: In many applications, it is important to make the discoveries more understandable by humans.
- User interaction / prior knowledge: Many current KDD methods and tools are not truly interactive and cannot easily incorporate prior knowledge about a problem except in simple ways.
- Integration with other systems: A stand-alone discovery system may not be very useful. Typical integration issues include integration with a DBMS, integration with spreadsheets and visualization tools, and accommodating real-time sensor readings.

2.7 Current DM Research Efforts and Related Works

In the domain of the health care there is relatively little work on what the data mining work best. Current efforts to turn information into knowledge in health care organizations are implementing this technology to help control costs and improve the efficiency of health care services.

Even though studies in Butajira have been conducted in a set of nine randomly selected (probability-proportional-to-size technique) rural kebeles (known as "peasants' associations") and one urban kebele (the Urban Dwellers' Association), only two research works done using DM [14, 64]. The rest research work conducted with statistical

software (like SPSS, EPI-Info) tools that mostly used to show verification (prove or disprove) instead to find unrecognized new knowledge from monthly visited to each household data. So far, three complete censuses of the population (in 1986, 1995, and 1999) have been done. The extent of similarity between the 1994 national census and the DSS database illustrates the quality of the continuous registration system. Currently, the surveillance interval is changing from monthly to quarterly. Custom-made software, based on the dBase system, is used to handle the data [11].

The study base is now well established and is being used for other more focused studies on essential health problems of the country, using qualitative, as well as quantitative, research methods. So far, research on childhood respiratory illnesses, other infectious diseases, reproductive health, and mental health has been conducted using the study-base infrastructure [5].

This work has contributed to human-resource development and research capacity-building at the Faculty of Medicine, Addis Ababa University.

Some research works for data mining in health care industry have been carried out but not yet fully investigated since the work was done at the time the amount of data as such not huge or large.

As cited by Dibaba [64] Abraham tried to identify determinant risk factors of HIV/ADIS infection and to find their association by using data mining application. The main objective of the researcher was to explore the potential factor application of DM to search for the major factors those results in HIV infection and transmission using voluntary counselling and testing (VCT) dataset in Addis Ababa. The researcher followed only a three steps methodology (Data collection, Data preparation and Model building and testing) implementing association rule discovery.

The researcher reported that 75% the data collected were from unmarried people showing that unmarried people are the major users of VCT services. Even though the size of the data was not mentioned, the researcher followed three step methodologies- data collection, data preparation and model building and testing. The researcher has disclosed

that people suspect and symptom is also associated to HIV negative results with evidences (36.33% support and 100% confidence).

The first study that was conducted at BRHP database using data mining application is by Anagaw [14] to predict child mortality patterns with only 1100 records. The researcher developed a model to support primary health care providers, policy makers and planners for identifying the major determinants of child mortality and to prevent and control child mortality.

The researcher has used the methodology suggested by Berry and Linoff [15]. 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.

The researcher obtained 93% and 95% accuracy with Neutral network and decision tree implementation, respectively. From his findings, the researcher reported that the results from both techniques indicate that data mining is an appropriate technology that should be employed to support child mortality prevention and control at the district of Butajira and even in general nationwide.

Recently, Dibaba [64] conducted data mining research to predict household health seeking patterns using BRHP dataset. The researcher aim was to develop a model that identifies risk factors and patterns of household health seeking behaviour at Butajira district.

A total of 60,446 records were used for the researcher's experiments with the implementation of J48 decision tree technique employed cross industry process for data mining (CRISP) methodology. The findings of the researcher indicated that predict household health seeking pattern using data mining techniques is possible. He found that the accuracy rate is 89.9017%. Even if the researcher tried to increase the accuracy of the model that was found, it was being unsuccessful and impossible. This implies that there was limited number of records to apply data mining technology (i.e. less than in hundred thousands).

Seen from the existing researches above, some methods have been already applied to information analysis; both domestic scholars regard the BRHP data as an important information source, and try to find out the hidden patterns through the use of different methods. However, most studies have stayed at quantitative features statistics with vital information, and they almost do not take into account the various bias effects of the data. While as for the utilization of vital information, it is basically on the level of qualitative analysis and information computational management, it cannot mine the knowledge rules automatically from the content of data.

Furthermore, the study of the vital statistics data in domestic programs is still at an initial stage, the actual disparity is greater compared with the developed countries, some of the researches aim at studying a certain vital statistics, yet there is no study on the vital statistics strength of the countries and it is only to evaluate the projects or analyze the country's overall situation. This type of analysis is more of vertical comparison with the health situation in Ethiopia over the past few years, as well as horizontal comparison with the vital quantity of the other countries, it cannot truly reflect the programs' value of vital information analysis.

However, the researcher of this study found only one related work on vital statistics data using data mining application that was conducted by Zhang et al. [65] at the state of California, USA. The researchers used a data mining tool called Cubist to build predictive models out of two million cases over a nine-year period. The objective of the study is to discover knowledge that can be used to gain insight into various aspects of mortality in California, to predict health issues related to the causes of death, to offer an aid to decision- or policy-making process, and to provide useful information services to the customers.

As the researchers report the results obtained in the study contain valuable new information. They wrote as the models produced by Cubist also contain surprising results that are not found in the official published reports. Most of those surprises represent valuable new information. Including marital status as an attribute during the mining process helped unearth valuable new information.

Considering these ideas, this study explore the use of data mining techniques for vital event information in our country, using classification method to analyse vital statistics data, in order to provide information support for the technical innovation activities of programs, it is an exploratory new idea. As to the researcher knowledge there is no research that attempt to apply data mining on vital statistics data in our country. Hence this study has a great contribution for discovering knowledge that can help for effective decision making and policy recommendations.

CHAPTER THREE

DM TOOL, TECHNIQUES AND ALGORITHMS

3.1 Introduction

As it is mentioned in the methodology section in chapter one, the problem that this research is going to address is a classification problem. Hence, it is important to explain the classification implementations for model building and experiments to be carried out in the data mining process, which also involve data mining tool selection and algorithms used for modeling.

To demonstrate real practicality in any data mining process, selecting the potential mining tool is important to understand clearly the techniques and algorithms to be implemented, and describe them specifically based on the tool used for the research work. In practical applications one has to decide which models and parameters may, be appropriate for diagnosis and prediction problems. An algorithm prove useful for a healthcare database may show not to be useful in a cooperate database. The tools used for data analysis are different: traditional statistical methods, neuronal nets (BrainMaker, NeuroShell), case-based reasoning (nearest neighbour), decision trees (See5/C5.0, Clementine), genetic methods (PolyAnalyst), machine learning algorithms and classifiers (WEKA), CRUISE, Discover*E, etc [66].

So, in this study the researcher chooses to use the Weka 3.6.2 software. The reason why this tool is specially selected is that it is the only toolkit that has gained widespread adoption and survived for an extended period of time and it is freely available for download (i.e. it is an open source software) and as well it offers many powerful features (sometimes not found in commercial data mining software), it has become one of the most widely used data mining systems. Other data mining and machine learning systems that have achieved this are individual systems, like Xelopes, not toolkits. Weka also became one of the favorite vehicles for data mining research and helped to advance it by making many powerful features available to all.

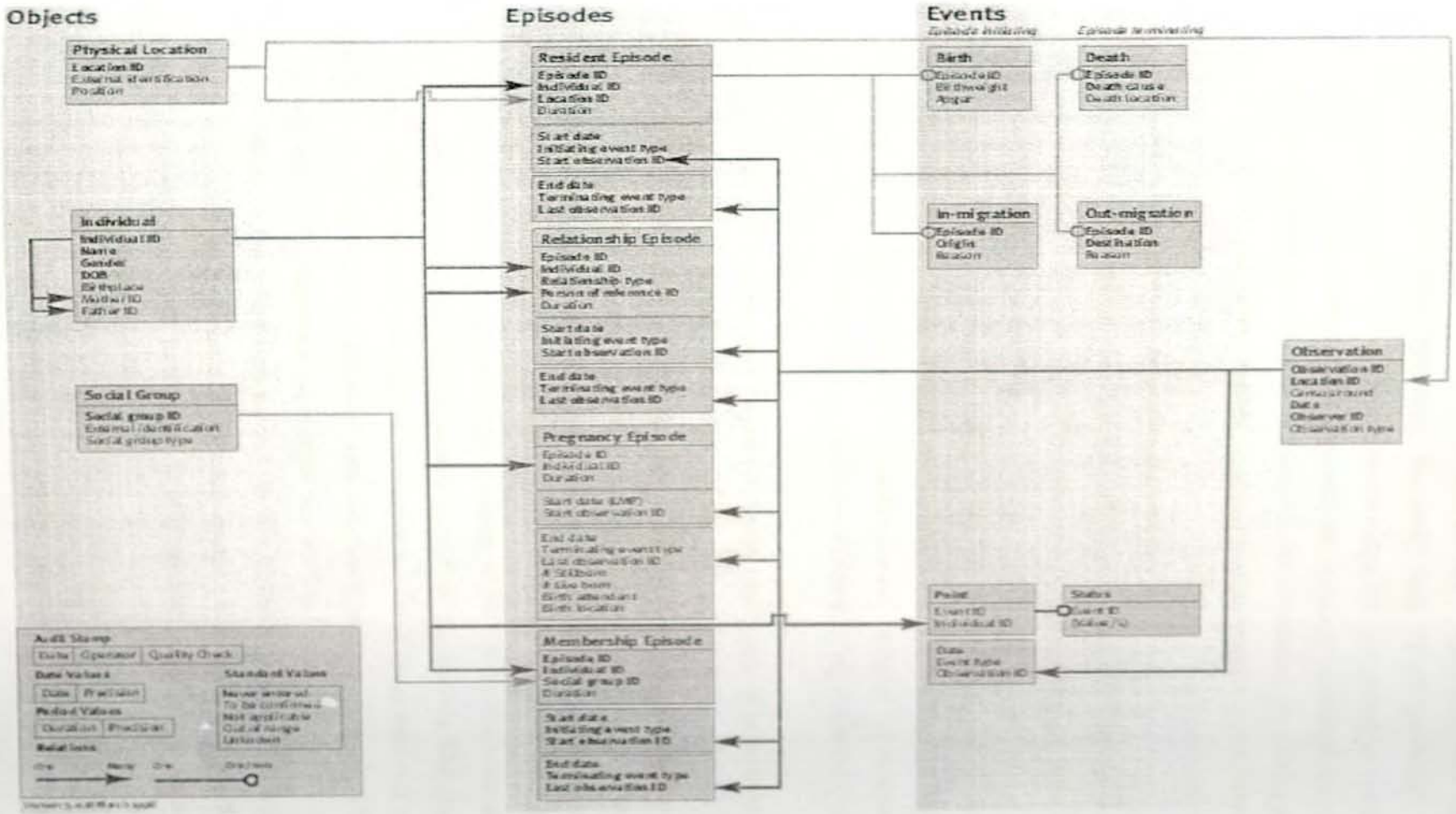


Figure 4.1 References Demographic Surveillance Data Model

- Individual — this entity contains a record for every individual who has ever resided in the study area. Optionally, this entity may record individuals whose residence in the study area has not been recorded but is required to complete a genealogy or relationship record. Records are uniquely identified through an individual ID value. Genealogical linkages can be established by storing the IDs of the individual's father and mother. This information (mother's and father's ID) can also be useful for identification purposes, especially where name and date of birth are not clearly defined, as is often the case in SSA.
- Social group — this entity stores information on a defined social group, such as a household. An individual is associated with one or more social groups, through one or more membership episodes.
- Observation — the observation entity stores the information that a particular physical location has been observed at a given time. This entity can also store information on the person making the observation and optional information, such as the census round. The observation entity is linked to all the events recorded during the observation.
- Events — the events entity may indicate a change in the state of an individual (for example, from resident to nonresident, in the case of an out-migration). Events that initiate and terminate a particular state of interest (for example, residency) are combined and recorded as an episode (for example, resident episode). These types of events are known as "paired events." Events that do not record the start or end of a particular state are known as "point events." The information common to all events (such as date of occurrence, type of event, and ID of the observation during which the event was recorded) is stored as part of the episode that this event initiates or ends (in the case of paired events) or as part of the point-event table (in the case of point events). Additional data associated with an event are stored in a separate entity. The following event types are noted in Figure 4.1:
    - Birth — this event type records all live births to residents (stillbirths are recorded as a pregnancy-outcome event). The event is linked to the resident episode it initiates — it also initiates social-group membership and relationship episodes.

- Death — this event type records all deaths of residents. A death event will terminate all open episodes belonging to the individual. The death-event record is linked to the resident episode that the event terminates and contains additional data, such as the location and cause of death.
- Relationship start — this event type records the start of a relationship of one individual to another. By convention, relationship events are linked to the female in cases of heterosexual relationships and to the younger individual in cases of same-sex relationships. In the case of caretaking relationships, the relationship events are linked to the person receiving care.
- Relationship end — this event type records the end of a relationship between two individuals.
- Membership start — this event type records the start of an individual's membership in a social group.
- Membership end — this event type records the end of an individual's membership in a social group.
- In-migration — an in-migration event initiates a new or changed physical location for an individual. It records the start of a new residence episode for an individual and can originate within or outside the study area. Additional data, such as origin, are usually stored in a separate entity linked to the episode via the episode ID.
- Out-migration — an out-migration event terminates a residence episode at a physical location for an individual. The destination of an out-migration can be within or outside the study area. Additional data, such as destination, are usually stored in a separate entity linked to the episode via the episode ID.
- Status observation — any number of optional events can be defined to record status information observed for individuals, such as socioeconomic, nutritional, educational, or immunization status. Repeated status observations make no assumptions about the value of observed attributes

during the observation interval, even if subsequent observations measure the same values.

- Episodes — As Figure 4.1 shows, episodes can occur to residents, relationships, pregnancies, and memberships in social groups:
    - Resident episode — a resident episode records the stay of an individual at a physical location. A resident episode can be initiated only by a DSS entry, a birth, or an in-migration event. It can be terminated only by a DSS exit, a death, or an out-migration event.
    - Relationship episode — a relationship episode records a time-dependent relationship, such as a marital union, between two individuals. The episode is started by a relationship-start event and concluded by a relationship-end event, a death, or a DSS exit. The relationship episode records the IDs of the two individuals involved in the relationship, but the events initiating and terminating the episode are linked to only one of the individuals, as described above.
    - Pregnancy episode — Pregnancy is recorded as an episode, with certain attributes recorded on the first observation of the pregnancy and others recorded when the outcome of the pregnancy is known. One lesson we have learned is that if you want to do a good job in child registration, you have to register pregnancies first. However, if a pregnancy is not observed, but only the outcome, the start of the pregnancy episode is still recorded as the date of the last menstrual period before the pregnancy. In this case the start and last observation IDs will point to the same observation instances. If a pregnancy is terminated by the woman's death or out-migration, the reason for termination is recorded as the terminating-event type, and the episode is concluded. In the normal course of events, the pregnancy outcome could be recorded in the terminating-event type as spontaneous abortion, induced abortion, normal delivery, assisted delivery, or caesarean section. The "birth location" field refers to the delivery environment (for example, the name of a hospital or clinic where the delivery took place).

- Membership episode — a membership episode records the membership of an individual in a particular social group. A membership episode can be initiated only by a DSS entry, a birth, or a membership start event. It can be terminated only by a DSS exit, a death, or a membership end event.

In summary, Figure 4.1 illustrates the entities and relationships of the reference data model. Mandatory fields and entities are displayed in bold type, whereas optional fields and entities are displayed in normal (nonbold) type.

4.1.2 Data collection and processing [5]

When the Butajira area was originally established as the DSS at 130 km far from Addis Ababa there were several reasons. First, it was considered beyond the direct influence of the municipal area but not too far from the university. Second, as it was in the mid-1980s civil war raged in northern Ethiopia. Hence a location to the south was preferred in the interests of long-term continuity. The area also offered a diversity of developmental, geographic, ethnic, and religious parameters within a fairly discrete area. As time passed, the extent of this diversity and its major consequences for many population parameters became increasingly apparent.

The initial census of the population in the selected villages was done in 1987 to obtain the baseline population and establish a system of DSS with continuous registration of vital and migratory events at household level. The total population was 28,780. Any adult member of the household >15 years old was eligible to respond in the monthly household interviews. These were carried out by a team of secondary-school graduate enumerators who were based in the kebeles. Each vital event was registered on a separate form at the household level. Basic demographic, social, housing-condition, and health-care-use characteristics were recorded for each household on its entry into the DSS and then during each re- enumeration.

As it happened, the first overall update of the 1987 census was not done until 1995, which was, in retrospect, too long. A further update round was then conducted in 1999. From the time of the 1987 census until 1999, continuous surveillance was carried out

during monthly visits to each household. However, in the light of experience, both here and elsewhere, quarterly household visits were phased in during 1999 and 2000.

4.1.3 Databases and records storage

An underlying principle for recording events in a DSS is that of a population at risk. Mortality, fertility, and migration rates are calculated by counting the number of deaths, births, or migrations occurring within a registered population exposed to the risk. For example, an individual who is not resident within the DSA is not considered at risk of dying within the area. Consequently, DSSs do not observe nonresident individuals or households and do not record their events. Before the surveillance interval is changing from monthly to quarterly, monthly visits to each household had provided the data.

As it is cited by Anagaw [14] Brahane et al stated that it was initially unrealistic to think of computerization of the data on-site for handling the data were limited at personal computers were still in early stages of development and very little experience of their use. So, data for the DSS were initially entered as text strings, but the DSS has, since 1994, used software based on the dBase IV platform. As developed for Butajira, this program includes procedures for automatic consistency checking and has more sophisticated facilities for data management and retrieval. The indigenous calendar used in Ethiopia runs behind the international calendar by 2809 days and has 13 months in a year, and this has presented serious obstacles to using proprietary packages for longitudinal data.

Data are currently entered in Butajira, which allows any inconsistent questionnaires to be immediately sent back to the field. This is a significant improvement over earlier practice, which was to centralize data operations in Addis Ababa.

The DSS operates as a dynamic open-cohort system. The individual person–years are aggregated to serve as denominators for calculation of various health and demographic indices. So far, three complete censuses of the population (in 1986, 1995, and 1999) have been done. The extent of similarity between the 1994 national census and the DSS database illustrates the quality of the continuous registration system.

Events registered by the BRHP are birth, death, marriage, new household, out-migration, immigration and internal move (migration within the BRHP DSS kebeles). Household and environmental variables are measured during the censuses. The study-base is now well established and is being utilized for other more focused studies on essential health problems of the country using qualitative as well as quantitative research methods. So far, research in the area of childhood respiratory illnesses, reproductive health, mental health and other infectious diseases have been conducted utilizing the study-base infrastructure.

Deaths and mortality

Deaths of all registered and eligible individuals are recorded, regardless of the place of death. It may be impossible to record the deaths of previously eligible individuals who then out-migrated. In this case, observation of their survival is censored at the time of migration. Information about the death of visitors to the DSA is sometimes collected, but it is only used in mortality estimates if a de facto population estimate is available for each day.

Underreporting of deaths is typically less of a problem than that of births, because a death is widely known and remembered. Exceptions are the deaths of young (and yet unregistered) infants, particularly prenatal deaths, if cultural beliefs or grief hinders reporting.

The DSSs collect more detailed information about deaths to establish the cause of death, generally through the so-called verbal autopsies (VAs).

The computer systems maintain standard DSS-processing operations:

- Data entry — Software allows for entry, deletion, and editing of the baseline and longitudinal data. Baseline household information includes the household location, individuals within the household, relationships between individuals, and familial social groups. Longitudinal information includes basic information on pregnancies and their outcomes, deaths, migrations in and out of the study area, marriages, and any other measures the investigators specify.

- Validation — Software checks for the logical consistency of data.

The site has also developed software for reporting outcomes and managing data:

- Reports and output — Routine software calculates and displays demographic rates and life tables and can compute age-specific and overall rates.
- Visitation register — Software prints the household-registration book, which is used by the fieldworkers to update and record information during household interviews.
- Utilities — this option is primarily used by the system administrator. It includes capabilities for adding new user IDs, setting interview-round information, and generating reconciliation reports to help track down unreported pregnancy outcomes and unmatched internal migrants.

4.1.4 Data quality assurance [5]

Data-quality-assurance mechanisms have been instituted at several points to ensure the integrity of the data. The most critical of these mechanisms is field supervision. Field supervisors daily supervise data collection and check each completed data form. They also make random visits to selected households each month, using a weekly distributed timetable. Research assistants supervise the data flow from the households to the computer system. They also check the data at the field level, for randomly selected households. Researchers work in the field to provide on-site technical assistance and guidance and check data quality.

4.1.5 Method of analysis at DSS [5]

Although the site reported data for longer periods, life tables were constructed in the standard fashion Preston et al [5]. nMx, the age-specific mortality rates for the age group x, x +n were calculated as the ratio of deaths, nDx, to person–years exposed, nPYx, in the same age group. When calculating nqx, the probability of dying in age group x, x+n, one assumes that the average age at death, nax, equals half of the age interval, except for ages <5 years. In the age intervals 0–<1 and 1–4 years, the values of nax are calculated using

the relationships developed by Coale and Demeny, based on the West model life-table system Preston et al [2]. The open age interval encompassing ages ≥85 years is closed in the usual way, by letting nL85 equal the ratio of l85 to ∞M85. Standard errors are calculated using formulae developed by Chiang (1984). For instance let's look at the following way of mortality data analysis.

A. Mortality data

To allow maximum flexibility in analysis, the site provide counts of deaths and person–years observed in standard 0 to 85+ age groups by sex for single years of observation for as many years of observation as possible. To adequately capture the variation in mortality over time, the data from the site is grouped into 3-year intervals, or as close to 3-year intervals as possible and practical

An individual's probability of dying depends primarily on sex, age, health, genetic endowment, and the environment, all of which determine the risk of falling victim to illness or accident [5]. The primary determinants of mortality interact in complex ways and depend in turn on a large and variable set of complex social determinants. As a result, it has not been possible to formulate a general, theory-driven model of individual risk of death.

B. Discussion

The data presented here are the first large compilation of high-quality data collected at intensively operated longitudinal field sites. In light of the general lack of high-quality information describing contemporary mortality, this is a unique and useful collection of data. The writer reported as the level of mortality varies considerably across the sites that have produced these data and all but one or two appear to have produced very reasonable age-specific mortality schedules. A great deal of additional analysis is applied to these data in the near future. The first extension of the basic description of the levels and age patterns of mortality presented here is the identification and thorough examination of the common underlying age patterns of mortality embodied in these data.

Furthermore, several Studies have been conducted in Butajira in a set of nine randomly selected rural kebeles (known as "peasants' associations") and one urban kebele (the Urban Dwellers' Association) implemented probability-proportional-to-size technique.

The intensity and diversity of the research activities have also required a wider participation of multidisciplinary researchers. The participating researchers have backgrounds in obstetrics, pediatrics, epidemiology and biostatistics, sociology, psychiatry, nursing, and public health. At present, more than 50 field staffs are working in the DSS [11].

The site manipulates and analyzes data with dBase, Epi-Info, and the cohort program, developed by Umeå University, which does person–year-based analyses of events in dynamic cohorts. For example, 5143 deaths and 15 667 births were registered in the area, from a total of 336 074 person–years of follow-up during 10 years of surveillance. Thus, based on the observed total number of deaths in this study base, the crude mortality rate is 15.3 per 1000 person–years. A total of 71 004 person–years has been observed among women 15–44 years old, representing 2367 reproductive lifetimes and hence an overall fertility of 6.6 births/woman. The maternal mortality ratio has been estimated using several methods and is believed to be around 600 per 100 000 live births [11].

Deaths among children <5 years old represent 48% of all mortality. Half of these deaths occurred during the first year of life, and 53% before 2 months. From the age-specific mortality rates we can estimate the cumulative mortality throughout life. Thus, among live births, an estimated 4.2% die during the first 2 months of life, 8.0% before 1 year, 16.6% before 5 years, 36% before 15 years, and 56% before 65 years. Substantial variations have occurred between areas with regard to under-five mortality, with rates ranging from 80 per 1000 person–years in the urban area to 219 per 1000 person–years in the lowlands. From the age-specific mortality rates, we estimate a current life expectancy at birth of 50.8 years — 49.3 years for males and 52.3 years for females. From these result obtained in the studies, national and international publications and scientific conferences had been the main routes to disseminate this information.

4.2 Data Pre-processing for Mining

At the preprocessing stage missing values can be filled in either by using Weka software or a separate tool (SPSS software) as explained later in the next section of the chapter. The training part of the cleaned data is first passed into the data mining tool where similarities in the patterns are extracted and model is created. Based on these patterns and rules obtained, classification of the testing data set takes place.

An objective of this study is to develop a model that can be used for description to gain insight into various aspects of mortality. We can implement these models to work well on a computer say a desktop or a laptop, but integrating the same tool on a handheld can be rather tricky.

Thus, instead of storing all the data and the data mining algorithms on the tool, we run the tool on desktop computers and sample only the run able data that the rule set on. We then input the data directly and the rule set can be run to provide the required answer.

Required each algorithm data to be submitted in a specified format, the generation of raw data into machine understandable format is called preprocessing [34]. Other steps that are performed during preprocessing are the transformation of the attributes in the database into a single scale and the replacement of all the missing values in the data.

Raw data can be stored in several formats, including text, Excel or other database types of files. Specifically in this study, data originally obtained in SPSS file format. So, it is necessary to convert this format to the system understandable format that the analysis is performed (i.e. Weka tool).

Having data already in a format understandable by algorithms can result in better time efficiency with respect to processing of the data. In most cases the rows represent a single case and columns represent the attributes that are present within this case. In some of the free databases that are available online most of them are in comma separated value (CSV) format. That is all the attributes are separated by commas and two commas

simultaneously stands for a missing data attribute. Sometimes when attributes are missing, instead of finding an empty space we may find a question mark in place of the missing attribute.

In the Weka tool, the data should be stored in the Attribute-Relation File Format (ARFF format) as the data type of the attributes must be declared. The system does not automatically classify the attribute as being real or categorical.

4.2.1 The raw data

The raw data usually has a great deal of noise, incompleteness, inconsistency, etc. Raw data cannot be used directly for processing, with the machine-learning algorithms. They first need to be preprocessed into machine understandable format. The database of Butajira Rural Health Program, 1987-2004 is considered as below to demonstrate preprocessing. The data types of the attributes with the raw data are given below.

Table 4.1 Attributes available in the eighteen years BRHP database

# Attribute	Type	Width
1. Year Reference (TTYREF)	Num	6
2. Peasant Association (PA) Code	Char	3
3. ENVIR	Char	1
4. HOUSENO	Char	5
5. ID (Compulsory)	Char	10
6. NAME	Char	20
7. Relationship (REL) code	Char	2
8. SEX	Char	1
9. Mother's ID (MID)	Char	10
10. Father's ID (FID)	Char	10
11. Marital Status (MARITAL)	Char	2
12. Serial foe Individual's Episode (SEREPI)	Num	2
13. Date of Birth (DBIRTH)	Date	8

14. Reason for Episode Starting (RSTART)	Char	2
15. Date of Episode Starting (DSTART)	Date	8
16. Date of Episode End (DEND)	Date	8
17. Reason for Episode Ending (REND)	Char	2
18. Date of Death (DDEATH)	Date	8
19. Date of Exposure (TIMEX)	Num	6
20. Cause of Death (CAUSE)	Char	1
21. Individuals' Religion (RELIG)	Char	2
22. Literacy during Episode (LITER)	Char	2
23. Educational Status (EDUCATION)	Char	2
24. Source of Water (SOURCEW)	Char	2
25. Type of Roof during Episode (ROOF)	Char	2
26. WINDOWS	Char	2
27. RADIUS	Num	2
28. ROOMS	Num	2
29. HOUSEOWN	Char	2
30. OXEN	Char	2
31. TIMAD	Num	2
32. LATITUDE	Num	8, 5
33. LONGITUDE	Num	8, 5
34. Distance to Butajira (DISTHOSP)	Num	2

This data set consisted of, a total of 87,092 records of individuals who were born and 9,699 records are about individual who died from January 1, 1987 to December 31, 2004 in all the ten villages of the BRHP study area. The following table shows the distribution of those birth and death throughout the eighteen periods in all the ten villages of the BRHP study area.

Table 4.2 Distribution of birth and death in the ten villages of the Butajira program

No.	PA(Peasant Assn.)	Environment	# of birth	# of death
1	Meskan	H	7,131	862
2	Bido	H	9,058	1,066
3	Dirama	H	8,785	1,534
4	Wrib	H	4,554	589
5	Yeteker	H	5,070	624
6	Bati	L	8,684	1,049
7	Dobena	L	10,752	1,184
8	Mjarda	L	5,656	726
9	Hobe	L	8,094	1,062
10	Buta04	U	19,308	1,003
		Total	87,092	9,699

Here we should note that all attributes are not important for the mining purpose in this research objective. So, finding a minimal set of attributes that preserve the class distribution is one of the main task in data mining preprocess.

4.2.2 Attribute selection

Selecting relevant features (attributes) in any data mining task is important for increasing the efficiency of the algorithm. As cited by Anagaw [14], Liu and Motoda (1998) wrote that "the abundance of potential features constitutes a serious obstacle to the efficiency of most learning algorithms.

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. Based on expert's opinion, only some of them are considered as relevant for the specific learning task to be undertaken in this research work. In this regard, the original data set consisted of a total of 34 attributes (columns) and 236,549 records (rows) from which the target data set for this research work has been selected.

In this research, the main focus is the pattern/relationship between different factors (attributes) and death (mortality) based on the research objectives. We therefore use those attributes related to the BRHP dataset as our classes and look for other attributes relevant

to these classes. After this step, we are able to filter out those irrelevant or less relevant attributes. Our data mining tasks thus become simpler and more accurate.

Hence the researcher decides 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. A total of 27 candidate attributes are selected as relevant features to build and test the required model. Selection of these relevant features is conducted in consultation with domain experts/specialists and the researcher's advisor who have good knowledge and experience on the data set of BRHP database.

In the database the attributes that are not contributed any information towards the machine intelligence in determining whether the individual has good aspect or not so the columns have been removed from all the cases within the database. The following table gives a brief description of the candidate attributes selected as relevant features to build and test the required models.

Table 4.3 Attributes selected from the original data file

No	Attribute Name	Description	Values
1	TYREF	A unique reference number given to each individual	A six digit serial number
2	PA	The name of the village in which the individual is registered	10 unique categorical values.
3	ENVIR	The climate of the individual's village	3 unique categorical values
4	SEX	The gender of the individual	2 symbolic values
5	REL	The relationship of the individual	7 unique categorical values
6	MARITAL	The marital status of the individual	8 unique categorical values
7	SEREPI	The serial no. of individual's episode	A two digit serial number
8	DBIRTH	The month/data/ year in which the individual was born	Date of the form DD/MM/YYYY
9	RSTART	The reason for episode starting	11 unique categorical values

10	DSTART	The month/data/ year in which episode starting	Date of the form DD/MM/YYYY
11	DEND	The month/data/ year in which episode ending	Date of the form DD/MM/YYYY
12	DDEATH	The month/data/ year in which the individual died	Date of the form DD/MM/YYYY
13	TIMEX	Days of exposure during episode	Up to four digit number
14	RELIG	Individual's religion	5 unique categorical values
15	LITER	Literacy during episode	5 unique categorical values
16	EDUCAT	Educational status during episode	5 unique categorical values
17	SOURCEW	Source of water during episode	7 unique categorical values
18	ROOF	Type of roof during episode	3 unique categorical values
19	WINDOWS	Windows in the house	3 unique categorical values
20	HOUSEOWN	The house ownership	5 unique categorical values
21	OXEN	The number of oxen owned by family	3 unique categorical values
22	LATI	The latitude of household	Continuous
23	LONGI	The longitude of household	Continuous
24	DISTHOSP	The distance to Butajira km	Continuous
25	RADIUS	The radius of circular house in meters	Up to two digit number
26	ROOMS	The number of rooms in house	A digit number
27	TIMAD	The number of timad of land owned by family	Up to two digit number

Following the successful selection of the required data set from the original eighteen years' data, the next important step considered by the researcher is filling up the missing values of the selected attributes.

4.2.3 Filling up missing and incomplete values

Sometimes there are attributes that are incomplete or missing. A common method of representing missing data, is inputting values that cannot be found in the data e.g.

represent missing data as "?". If an attribute is empty usually one may think that the case is less useful than the rest of the cases in the data set. This is not true as each of the other attributes contributes useful information towards the set of attribute category. When there are missing values, instead of leaving them as missing, there are a number of methods that can be used for filling these missing attributes.

Having efficient methods to fill up missing values extends the applicability in terms of accuracy for many data mining methods. The accuracy of the tool is increased and with a larger training set better rules and decision trees can be developed which contributes towards better classification of the data.

The most common method of filling the attributes quickly and without too much computation is to replace all the missing values with the arithmetic mean for numeric data, the median for ordinal data or the mode for nominal data with respect to that attribute. The other methods are to run a clustering algorithm and replace the missing attributes with the attributes of cases that appear close in an n-dimensional space. In this case the researcher apply to use the statistical tool for replacing the missing values by series mean method since the original dataset is obtained in SPSS file format and all the attribute missed their values are numeric. That is, there is no need to use any other method including the above mentioned here for filling the missing values.

Table 4.4 Handled missing values

No.	Attribute Name	No. of missing values	Attribute Type	Substituted Values	Method used
1	Radius	77615(33%)	Scale(Numerical)	3.4	Mean
2	Rooms	6498(3%)	Scale(Numerical)	1.8	Mean
3	Timad	39263(17%)	Scale(Numerical)	2.9	Mean

4.2.4 Data decoding and attribute transformation

It is often necessary to transform values of attribute to another for making the attributes useful in the prediction modeling process. Particularly, 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 [65].

In the original data set created for this data mining task, the values of some attributes are represented using numerical codes. Such numerical codes that 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:

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

Table 4.5 Actual Values of each PA code and the original numeric code.

Original "PA" code	New value represented
005	Meskan
007	Bati
008	Dobena
011	Bido
04B	Dirama
06A	Yeteker
06B	Wrib
09A	Mjarda
09B	Hobe
K04	Buta04

In addition the following data transformation and reformatting operations are employed 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.

Creating the status attribute: The attribute "**status**" is not included in the original data set instead date of death as an attribute. Using this newly created attribute as one dependent variable can help to classify individuals into different groups. This classification would help to predict the likelihood that a given individual would die or survive. So, the status attribute is created from the date of death attribute.

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

Using Weka tool, three numeric attributes (Age, Ystart, and Yend) are discritized. For example, the created age attribute is discritizzed by using bin method in Weka. The number of bin is adjusted to be nine. Similary the rest two numeric attribute is also discritized in similar fashion to make modeling process easy for the selected techniques and algorithms in classification purpose.

In order to build a model that can be used to predict mortality pattern in the study area based upon environmental, parental, and health related factors, a data set consisting of both classes' of status (Dead, and Alive) is created and prepared for analysis since the classifier should be known for the predictor.

After the successful preparation of the required data set from the original eighteen years' data, the next important issue considered by the researcher is importing the selected, created and reformatted data set, which was in SPSS document format into Weka software understandable format.

4.2.5 Machine understandable format in Weka

Although the original data set is organized in rows and columns using SPSS statistical software, it is not possible to import this data set as it is into the data mining software that are used in this research work. Since most data mining tools can use data in the comma separated value (CSV) format for running the machine intelligent algorithms. The researcher decided to import the original data set into CSV format, which will then be imported to Weka software. As a result of this attempt, the researcher has successfully imported the selected data set into attribute relation file format (ARFF).

The data that is used for Weka should be made into the following format shown in the table below and the file should have the extension dot ARFF (.arff). For example, in order to perform decision tree classification technique, which is a supervised learning that required predefined class to train and build models identifying the relevant attribute, is important. The last attribute where the classification of the individual is done is made into a categorical format. That is, the classification attribute 'status' takes string values 'Dead' when death occur and 'Alive' when death not occur.

Table 4.6 Sample Weka system understandable ARFF format for BRHP dataset

```
@relation 'BRHPDataset'

@attribute PA {Mmeskan,Buta04,Bati,Wrib,Dobena,Mjarda,Hobe,Bido,Dirama,Yeteker}
@attribute ENVIR {H,U,L}
@attribute REL
{NE,CH,RE,SP,GP,UK,NR}
@attribute SEX {M,F}
@attribute MARITAL
{UK,MO,TY,NM,WI,PO,DI,SE}
@attribute AGE numeric
@attribute RSTART {IN,MU,XX,BI,ST,CM,MO,LI,RO,WA}
@attribute YSTART numeric
@attribute YEND numeric
@attribute TIMEX numeric
@attribute RELIG {MU,OC,UK,CH,OT}
@attribute LITER {IL,LI,TY,UK,RE}
@attribute EDUCAT {UK,NO,PR,SE}
```

ribute SOURCEW {RI,WU,PI,LA,UK,WP,OT}
ribute ROOF {TH,UK,CO}
ribute WINDOWS {NO,YE,UK}
ribute RADIUS numeric
ribute ROOMS numeric
ribute HOUSEOWN {OW,UK,OT,KE,RE}
ribute OXEN {UK,YE,NO}
ribute TIMAD numeric
ribute LATI numeric
ribute LONGI numeric
ribute DISTHOSP numeric
ribute STATUS {Alive,Dead}

ata
001,Mmeskan,H,HE,M,UK,1,51.28,IN,1994,1995,365,MU,IL,UK,RI,TH,NO,4,?,OW,UK,?,8.09588,
7672,3.5,Alive
446,Yeteker,H,HE,M,UK,1,93.69,ST,1987,1991,1669,MU,IL,UK,RI,TH,NO,?,1,OW,UK,4,8.05935,
0893,10.4,Dead

In this research work, the essential file for any Weka application, which describes the attributes selected to build the models and their classes, is created (See Annex A). This file contains the cases that are analyzed in order to produce the classifier. 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 example, the first three cases from the data file prepared for this research work looks as above:

A row represents one individual's case with values of attributes mentioned above separated by a comma. 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.

Weka can use another 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 training file.

In this research work, a separate test file is not prepared. Rather, the performance of the classifier is evaluated by using the most common test option, cross validation. Thus, by

invoking these options, training and testing samples cases are randomly selected from the file. Although another more optional classifier evaluation is available used by Weka for the cases file and as such is not used for this research work.

4.2.6 Balancing the target attribute with SMOTE

Once the dataset imported into the Weka system, it is important took the target attribute whether it has a balanced classes or not. According to Larose [16] if one class of the target attribute has much lower relative frequency than the other class, balancing these classes is recommended. Because the classification model could simply predict for some class that have more relative frequency to all operations and achieve 99% accuracy.

In this regard, the researcher identify similar problem mentioned above in BRHP dataset in which 56,424 (85%) are alive while 9,699 (15%) died individuals after sampling of 66,123 cases from the total of 236,549 BRHP dataset is performed. Although there are different approaches to solve the imbalance [17]: under-sampling of the majority class and/or over-sampling of the minority class, over-sampling of the minority class has been proposed by creating "synthetic" examples. This helps us as a good means of increasing the sensitivity of a classifier to the minority class. In this situation, the researcher tries to balance the cases for the target attribute by using Synthetic Minority Oversampling Techniques (SMOTE) which is provided from Weka supervised instance filter option and considers a total of 95,220 (56,424 alive and 38,796 dead).

Following this chapter we are running experiment for the selected mining tool, technique and algorithm using the prepared dataset to build the models and identify the best one for the research objective.

CHAPTER FIVE

EXPERIMENTATIONS AND RESULT DISCUSSIONS

5.1 System Architecture

In this research vital statistics data is mined to extract patterns related to death issues with the help of Weka tool. Using the JDM process, a system is designed that courses the classification. The system is trained using vital statistical data. We follow cross-validation to evaluate the system. The experiment is run on data set that consists of 25 attributes and 95,220 records after SMOTE. Based on these sampled dataset, it is ample enough for understanding the different stages that are used in various data mining algorithms.

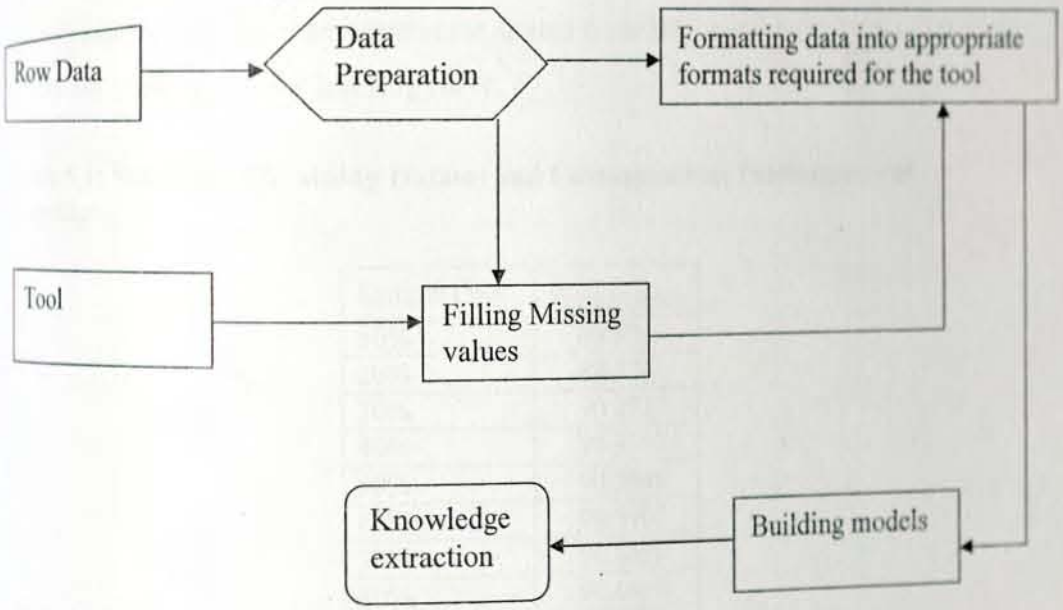


Figure 5.1 an overview of system Architecture

The above Figure 5.1 shows the different components of the system. In order to carry out the data mining process, the flow of the system starts with the collection of raw data. This data is first preprocessed and converted into formats understood by the Weka tool that are used in the DM process as we discussed in chapter three of this paper.

5.1.1 Estimating the Error Rate of the Learned Models

Before we run the classification analysis to make prediction, it is important to check the appropriateness of dataset for selecting certain validation method for the available dataset. Because after collecting the dataset learning algorithm is applied in independently. Thus the algorithm has used the training set to train the classifier, accuracy is estimated on the test set. To get an idea of how much data is needed in practice, each set of experiments is run on several different training set sizes.

To see the appropriateness of the dataset preprocessed and ready for model building, the researcher tried to observe its learning curve. From the graph on the learning curve, the point at which the learning curve converges might show the minimum of sample dataset to be used for training purpose so that the rest is planned for testing purpose.

For this research purpose the experiment started from 10% gone through up to 100% of the dataset as shown on the learning curve.

Table 5.1: Samples of Training Dataset and Corresponding Performance of Classifier

Sample (%)	Performance
10%	89.8040
20%	89.9754
30%	90.2717
40%	90.3503
50%	90.5046
60%	90.5967
70%	90.6090
80%	90.6690
90%	90.6788
100%	90.7520

Figure 5.2 shows performance on the test set as a function of the size of the training set. The basic training set exhibits satisfactory performance with error rates being low across all learning tasks. In general, the learning curve always has lower error rates across all four tasks on the larger data sets sizes.

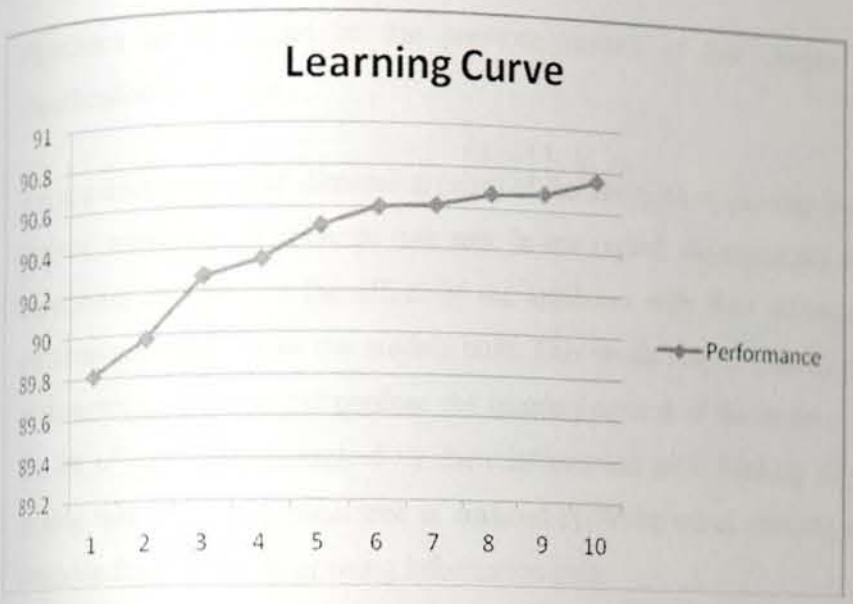


Figure 5.2: Learning Curve for Training Dataset

On the smaller data sets, the learning curve sometimes achieved lower error rates. The lowest error rates for all training tasks are achieved by using a data set size of 28,566 examples.

As we can see from the learning curve, the saturation point goes sharply up as the sample size increases. To minimize the effect of the shortage of the dataset used, the researcher select to use 10-fold cross validation for 95,220 cases for which it is sampled from the whole 236,549 datasets of BRHP database.

5.2 Experimentations

The other objective of this research work is to predict the mortality whether the individual is 'Dead' or 'Alive' from data obtained from the BRHP database in order to extract patterns that shows the risk for death.

The BRHP database consists of more than 236,000 cases. A section from this database is used for the testing stage and the rest for training. It is always a good practice to have a larger set of data for training than for testing [9]. In this case we use the data set into 9 folds training cases and the rest 1 folds of the cases for testing the data mining

algorithms as described in the previous section of this chapter to conduct the classification technique.

It is apparent that as the dimension (size) of the tree goes on growing it turns into solid to analyze, interpret and generate rule sets. In this regard, the researcher has carried out an experiment to looks at the effect of the attributes with their information gain on the performance accuracy of the models built. This would help to identify that the irrelevant or diverting attributes that confuse the learning process of the models. Table 5.2 shows the list of attributes as ranked by their information gain. Ranking the attributes to the mining task of the decision tree is realized by Weka select attribute option tab that is available for ranking filter using information gain.

Table 5.2 the ranked attributes with their information gain

No.	Attributes Name	Information Gain	Deviation from the maximum (i.e Max=0.10825)
1	ROOF	0.10825	0
2	WINDOWS	0.09871	0.00954
3	TIMAD	0.09819	0.01006
4	ROOMS	0.09131	0.01694
5	RADIUS	0.08487	0.02338
6	DISTHOSP	0.07132	0.03693
7	ENVIR	0.06849	0.03976
8	HOUSEOWN	0.05925	0.049
9	RELIG	0.05658	0.05167
10	YEND	0.05148	0.05677
11	MARITAL	0.04992	0.05833
12	PA	0.04911	0.05914
13	TIMEX	0.04628	0.06197
14	RSTART	0.04051	0.06774
15	SOURCEW	0.04027	0.06798

16	REL	0.03852	0.06973
17	YSTART	0.03115	0.0771
18	LATI	0.02869	0.07956
19	AGE	0.02223	0.08602
20	LONGI	0.02132	0.08693
21	EDUCAT	0.02079	0.08746
22	OXEN	0.02057	0.08768
23	LITER	0.01545	0.0928
24	SEX	0.00683	0.10142

Being familiar with attributes' importance to the data mining task as portrayed above. According to the information obtained from the above table, the researcher preferred to run the experiment using all attributes that are selected in the attribute analysis section of chapter three.

So, as we can see from the result of attribute selection using entropy based information gain method of weka, twenty-four attributes of the dataset are determining for predicting the STATUS variable that is created for prediction. Seven attributes are separated from these attributes having high importance for model building by identifying their little disorderliness (less discriminative). Again, the top two attributes are also took apart from these significant attributes due to having potential embark to classify the model perfectly as the class attribute related to their very high disorderliness. Hence the researcher performs the experiment using these best attributes and the identified 10- fold cross validation for eight scenarios are conducted.

The details of the decision tree used in WEKA are explained in detail in previous chapter three. For the decision tree to be created, algorithms are required to be executed from the training data following the test data. Once the trees are extracted and selected the best one, the rule is created based on the tree and the association between the attributes. The decision trees with respect to the algorithm described in chapter three for BRHP database research is shown. Classification on the test data is done based on the decision tree that is created.

All the classification techniques have similar screens. The bottom right section of the screen displays the classifier output. The classifier outputs results based on the majority class, that is, the outcome of the experiment which is always the class with maximum number of cases. This is considered the study case in this research and also takes the least computation time. Classification of data and the confusion matrix is displayed in the classifier output screen below the decision tree.

Experiments are performed with various method parameters, mainly changing types of decision tree. The J48 classifier window enables us to switch the parameters to build different decision tree scenarios.

It is also important to describe the J48 classifier parameters those allow us for intelligently adjusting them. Bellow in Table 5.3 J48 classifier Parameter Options are described which is taken from Weka Manual. Often only repeated experiments and familiarity with the data tease out the best set of options as shown below.

Table 5.3: Description of J48 classifier Parameter Options in Weka

Parameter Options	Description
binarySplits	Whether to use binary splits on nominal attributes when building the trees.
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning).
debug	If set to true, classifier may output additional info to the console.
minNumObj	The minimum number of instances per leaf.
numFolds	Determines the amount of data used for reduced- error pruning. One fold is used for pruning, the rest for growing the tree.
reducedErrorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning.
saveInstanceData	Whether to save the training data for

	visualization.
seed	The seed used for randomizing the data when reduced-error pruning is used.
subtreeRaising	Whether to consider the subtree raising operation when pruning.
unpruned	Whether pruning is performed.
useLaplace	Whether counts at leaves are smoothed based on Laplace

Based on Weka data mining tool implemented J48 classifier with different parameter, the researcher employs J48 algorithm for applying decision tree classification model on the BRHP dataset preprocessed as in the previous chapter.

There are eight scenarios that are experimented for decision tree classification in this research. These scenarios are analyzed to compare them to each other in terms of different performance matrices values, accuracies, number of leaves, and size of tree generated, ROC curves and execution time.

The scenarios for decision tree classification that are experimented in this research are as listed below.

Scenario #1: General Decision Tree pruned without missing value replacement before SMOTE

Scenario #2: Binary Decision Tree pruned without missing value replacement but before SMOTE

Scenario #3: General Decision Tree pruned with missing value replacement before SMOTE

Scenario #4: Binary Decision Tree pruned with missing value replacement before SMOTE

Scenario #5: General Decision Tree pruned without missing value replacement after SMOTE

Scenario #6: Binary Decision Tree pruned without missing value replacement after SMOTE

Scenario #7: General Decision Tree pruned with missing value replacement after SMOTE

Scenario #8: Binary Decision Tree pruned with missing value replacement after SMOTE

From these experimental scenarios, we obtain the models result to compare them each other and finally we find out the outperforming model based on the criteria of evaluation.

The data mining experiments are run and the outputs are obtained by the Weka tool is summarized in table 5.4. This study has attempted to look at the effects of mortality in a rural context encompassing as many confounding factors or variables as possible.

Table 5.4 Summary of measures of performance and accuracy of the models

Experimentation	Before SMOTE				After SMOTE			
	Without replacing missing value		With replacing missing value		Without replacing missing value		With replacing missing value	
Scenarios#	1	2	3	4	5	6	7	8
Tree Type	Pruned General	Pruned Binary	Pruned General	Pruned Binary	Pruned General	Pruned Binary	Pruned General	Pruned Binary
# of leaves	167	71	167	71	321	107	338	112
Size of Tree	208	141	208	141	401	213	424	223
Time(Sec)	7.57	31.62	8.2	33.38	10.42	38.01	12.12	37.29
CCI	90.10	89.51	90.10	89.50	90.23	89.50	90.30	89.50
AVG TPR	0.901	0.895	0.901	0.895	0.902	0.895	0.903	0.895
AVG FPR	0.423	0.442	0.423	0.441	0.113	0.118	0.112	0.118
Precision	0.893	0.886	0.893	0.885	0.902	0.895	0.903	0.895
Recall	0.901	0.895	0.901	0.895	0.902	0.895	0.903	0.895
ROC Area	0.885	0.892	0.885	0.892	0.950	0.941	0.951	0.941
Sensitivity	0.968	0.965	0.968	0.964	0.936	0.936	0.935	0.924
Specificity	0.509	0.488	0.510	0.490	0.853	0.853	0.856	0.853
F-Measure	0.893	0.887	0.893	0.887	0.902	0.895	0.903	0.895

5.3 Result Discussions

From the eight scenarios, results have shown that most parameters implemented have outperformed the decision tree that are created by changing parameters from the weka Generic ObjectEditor window for building different decision trees using J48 algorithm. An added advantage of J48 based algorithm is that there are more cases present in category "Alive" than "Dead" to produce interpretability for the health practitioners and may help in both the validation of the method and in developing further knowledge of the problem.

One of the steps in the knowledge discovery process is to evaluate the performance of the system in terms of how correctly the model classifies records in to different labeled classes. Sometimes, the actual class and the classifier decisions may differ in predicting a record to a certain class label. Hence the researcher is concerned to examine the decision trees built using all the eight scenarios to compare its efficiency experimented so far. As we can see from the table 6.4, the accuracy measures are obtained by altering parameters set in the J48 algorithm.

From this, we can say that scenario #7 has the best accuracy performance which is 90.30. The mean absolute error which measures the error between the actual and the predicted value is also lower at this best scenario, 0.151.

The performance of scenario #5, in all aspects is next to the performance experiment seven. The first scenario takes smallest time of all the scenarios. The table shows that the maximum number of tree and biggest size of the tree is experiment 7 and the least is experiment 2. The researcher also prefers that it good to visualize through line graph the performance accuracy's used to analyze and evaluate the models to select the best one for further discussion and easy understanding. Figure 5.3 compares the performance of the models in correctly classifying instances. The minimum performance observed on experiment (scenario) #2, 4, 6 and 8 with the maximum for the scenario # 7.

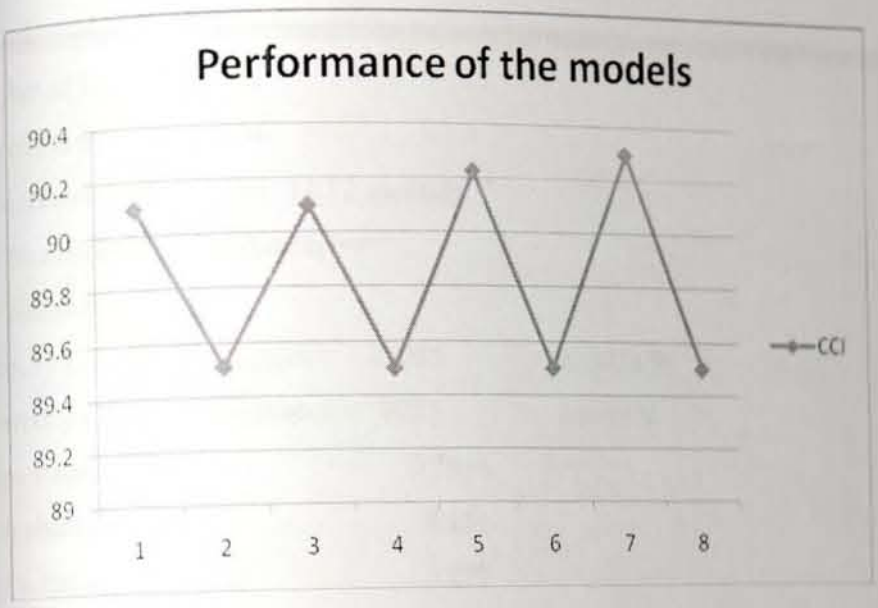


Figure 5.3 Comparison by performance for all models

The data mining output represents a significant advance in the type of analytical tools currently available although there are limitations to its capability. One limitation is that although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship.

So to be successful, data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Due to lack of space and time constraint from these eight models, only the best selected model is discussed in this paper with most meaningful relations to humans and with best classification accuracy at the same time. That is the generalized decision tree pruned with missing value replacement after SMOTE.

In this regard, the accuracy in terms of percentage is obtained from the classifier output which is 90.30 as shown in Figure 5.4. This accuracy is obtained with the adjusted parameters set in the J48 algorithm.

Number of Leaves : 338

Size of the tree : 424

Time taken to build model: 12.12 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	85985	90.3014 %
Incorrectly Classified Instances	9235	9.6986 %
Kappa statistic	0.7976	
Mean absolute error	0.151	
Root mean squared error	0.277	
Relative absolute error	31.2624 %	
Root relative squared error	56.3766 %	
Total Number of Instances	95220	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.935	0.144	0.904	0.935	0.92	0.951	Alive
	0.856	0.065	0.901	0.856	0.878	0.951	Dead
Avg.(Wt)	0.903	0.112	0.903	0.903	0.903	0.951	

=== Confusion Matrix ===

a	b	<-- classified as
52779	3645	a = Alive
5590	33206	b = Dead

Figure 5.4 Classifier output based on J48 decision trees.

The classifier predicts the records in to a certain class as there are similar attributes that lie in the same class boundary. However, the attribute that determines the class boundary of the given record is suppressed due to the data-driven (attribute similarity) trend applied by the classifier. Such kind of attributes similarities that tend in the incorrect classification of records. This shows that the record that is labeled by the actual class may

be labeled by the classifier to other class. This kind of phenomena often reduces the performance of the system.

Hence as the above figure indicates that out of the unseen instances (test data) supplied ,this model, using the best scenario can correctly classify 90.30% of them in their proper class, while 9.70% of them are misclassified into different classes.

Table 5.5 Sample of instances that show the actual and predicted class difference

	Instance number	Values of the Attribute							
		95216	95208	95219	95192	7	67	51	9
Name of Attribute selected for this study	PA	Hobe	Hobe	Mmeskan	Mmeskan	Dobena	Dobena	Mmeskan	Mmeskan
	ENVIR	L	L	H	H	L	L	H	H
	REL	CH	CH	CH	HE	HE	HE	CH	CH
	SEX	F	M	M	M	M	M	F	F
	MARITAL	TY	TY	TY	MO	MO	MO	TY	TY
	AGE	0-10.5	0-10.5	10.5-21	31.5-42	63-73.5	63-73.5	0-10.5	0-10.5
	RSTART	MU	MU	XX	XX	XX	XX	MU	XX
	YSTART	2002-2004	1995-1997	1993-1995	1995-1997	1995-1997	1995-1997	1997-1998	1995-1997
	YEND	2002-2004	1998-2000	1993-1995	1995-1997	1997-1998	1997-1998	1998-2000	1998-2000
	TIMEX	64	874	510	516	797	797	317	1438
	RELIG	MU	MU	MU	MU	OC	OC	MU	MU
	LITER	TY	TY	TY	LI	UK	UK	TY	TY
	EDUCAT	NO	NO	NO	NO	NO	NO	NO	NO
	SOURCEW	WP	WU	RI	RI	RI	RI	WP	RI
	ROOF	TH	TH	TH	TH	TH	TH	TH	TH
	WINDOWS	NO	NO	NO	NO	NO	NO	NO	NO
	RADIUS	3.0	3.6	3.1	5.0	3.0	3.0	3.0	4.0
	ROOMS	1	1	2	1	1	1	1	1
	HOUSEOWN	OW	OW	OW	OW	OW	OW	OW	OW
	OXEN	NO	NO	YE	YE	NO	NO	YE	YE
	TIMAD	1	2	3	3	5	5	2	4
	LATI	8.03	8.04	8.08	8.09	8.10	8.10	8.09	8.10
	LONGI	38.47	38.47	38.34	38.35	38.44	38.44	38.36	38.38
	DISTHOSP	11.5	14.3	7.1	5.2	7.5	7.5	4.7	3.5
STATUS	Actual	Dead	Dead	Dead	Dead	Alive	Alive	Alive	Alive
	Predicted	Alive	Dead	Dead	Alive	Dead	Alive	Alive	Dead

As we can perceive here the result of the classifier vary in classifying a certain instances differently that is in the actual class. From the incorrectly classified instances, the researcher try to identify for which instances the classifier misclassify the instances for being reduced the performance accuracy measure as depicted in Table 5.5.

The result obtained in the study contains valuable new information. This result conveys some interesting findings. Hence we are able to understand the problem of misclassification under the considered dataset. As shown in figure 5.5, the live status of individual as it is indicated by instance number 95208 and 95219 are actually 'Dead' whereas instance number 95216 and 95192 are classified as 'Alive' by the classifier even their actual class is 'Dead'. The researcher observes in similar fashion of the other class (in this case 'Alive') and clearly identify the problem of misclassification as we do in the 'Dead' class and presented in Table 5.5 above. In all the cases the misclassification occurs from the misjudging of a single attribute values calculating the similarity of the other attributes disregarding their difference as the principal predictive values.

Now, let us discuss at the confusion matrix that contains information about actual and predicted classifications done by a classifier. The data in a confusion matrix can be used to evaluate performance of classifiers.

A confusion matrix for the adjusted parameter is observed in addition to the error rates over various data set sizes. The confusion matrix is a matrix showing the predicted and actual class. In this experiment we have two types of classes. That means the outcome of the experiment is either the individuals is 'Dead' or 'Alive'. Thus a confusion matrix with two classification look like the table given above.

In Table 5.4 it is possible to notice that the table (matrix) shown the columns represent the predicted result and the rows represent the true or actual result. Hence from Table 5.4 the number 52779 and 33206 indicates the number of individuals where the actual and predicted values are similar. The number 3645 represents the number of individuals where the actual outcome is 'Alive' but is classified as being 'Dead' by the WEKA Decision tree classifier. Similarly the number 5590 represents the number of individuals

where the actual outcome is 'Dead' but wrongly classified as being 'Alive' as shown in the Table 6.4.

The number of true positives in this confusion matrix is 52779 records. Those records which are predicted as 'Alive' class by the classifier and also happened true by when tested on the test data are (True Positives). The number of the records which are classified to the "Dead' class by the classifier and they are actually False as tested on the test data (True Negative) is 33206. The sum of these values (equals to 85985) gives us the correctly classified number of cases even if we provide 95,220 as the total datasets (cases) to the experiment.

As we can also see from the Table 5.4, sensitivity is a bit high for this selected model while specificity is next to sensitivity but high across even in all other models. Sensitivity is the ability of the classifier to identify true positives (in this case the actual Alive), which is equal to the recall of the classifier whereas specificity of the classifier is the ability of the classifier to identify the true negatives (Actual Dead in this case).

Precision and recall is also a good performance measure for the model built. The precision of this model for the 'Alive' class is a bit higher than the precision for the 'Dead' class. The ROC area of all of the models is also above 0.5 is moderately good. The ROC is a simple graphical plot of the sensitivity, or true positive rate, vs. false positive rate (1 – specificity or 1 - true negative rate), for a binary classifier system as its discrimination threshold is varied. The ROC can also be represented equivalently by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate). It is also known as a Relative Operating Characteristic curve, because it is a comparison of two operating characteristics (TPR & FPR) as the criterion changes.

5.3.1 Generating rules from the decision tree

The size of the tree given by the number of nodes and the number of leaves (Classification nodes) that is present in the tree are shown in Figure 5.4. From this

tree it is seen that only 15 of the attributes are required to create the tree which means the predictor attributes that are used for classification of the dataset. The Partial decision tree constructed is presented in Annex B.

From the decision tree created it is possible to understand the meaning of the patterns and generate rules from that decision tree. Because rules are more accurate predictors of a given class than decision trees developed. Rules are generally easier to understand than trees since each rule describes a specific context associated with a class. Furthermore, a rule set generated from a tree usually has fewer rules than the tree has leaves.

We take a close look at the decision tree and find that the top level of tree is the attribute "ROOF". Therefore we can infer that this attribute is most discernable to the target value we want to predict. The classifier construct by using 15 attributes has an accuracy of 90.30%. This classifier has used some attributes to construct rules and provided the class predicted by the rule.

By default rules are ordered by class and sub-ordered by confidence. The number of instances for each label/class is given at the leave. 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. Follows the name of the majority class, the number of instances for the majority class/ number of the minority class is available in brackets. From this, it is possible to calculate the likelihood predictability of the majority class from the numbers of instances. For example, Dead (2401.0/35.0) is the first leaf of the decision tree in the generalized decision tree with pruning using some attributes model, in which 'Alive' value is 2401/2436= 0.99 likelihood predictability of the majority class. Similarly, we can calculate level of confidence/likelihood predictability of the outcome for the other majority classes. The following are a few of the patterns/rules which are discovered between status and other attributes for which their predictability of the majority class is relatively high compared to the rest.

Rule1: If ROOF is 'TH' (Thached) and TIMEX is '<= 1380.98' and if RSTART = is 'IN' (in migration) in which the individual is registered for the study is then the status is predicted as 'Alive' (will survive) highly likely.

Rule2: However, If ROOF is 'TH' (Thatched) and TIMEX is '<= 1380.978386' and if RSTART is 'MU' (multiple change) and YEND is within '1995.5-1997') the individual involves at the study area is then their status is predicted as 'Dead' (will not survive).

Rule3: However, If ROOF is 'TH' (Thatched) and TIMEX is '<= 1380.978386' and if RSTART = is 'XX' (join between 1995131 and 20050101) and RELIG is 'MU' (Muslim), and if MARITAL is 'TY' (too young) then individuals in the study are highly likely Dead (will not survive).

Rule4: Those individual However, If ROOF is 'TH' (Thatched) and TIMEX is '<= 1380.978386' and if RSTART = is 'XX' (join between 1995131 and 20050101) and RELIG is 'OC' (Orthodex), and if AGE is '<=10.5' and '52.5-63' is then the individuals have high chance not to survive for the study.

Rule5: If ROOF is 'TH' (Thatched) and TIMEX is '<= 1380.978386' and if RSTART = is 'XX' (join between 1995131 and 20050101) and RELIG is 'MU' (Muslim), and if MARITAL is 'NM' (never married) and EDUCAT is 'UK', if ENVIR is ' H' (highland) then individuals in the study are highly likely Dead (will not survive) whereas if ENVIR is ' L' (lowland) then individuals will survive in the study area.

As it is observed from rules 1 to 5 above, mortality is associated with types of roof ,reason start the episode, days exposed for episode, educational background, age of the individuals, individuals religion, year end the episode, and even with inmigration.

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 is evaluated based up on comments given by domain experts and reports of previous research works.

If it is possible it is better to experiment these variables using association rule discovery. However, the researcher is restricted to perform further study due to time constraint for this research work. Hence the researcher has recommended as further experiment in the next chapter of this thesis for concerned bodies or researchers to validate the relevancy of the attributes.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

In this research an attempt is made to investigate the possible application of data Mining technology in developing a model that can support for identifying the patterns of vital event. Result of the research is an important milestone in placing appropriate interventions rules and regulations that can minimize the burden of public health problems. Furthermore, the very interesting and unknown characteristic of the mined result is igniting further researches to be embarked by public health and IT professionals.

Machine intelligence algorithms are improving as the number of data mining tools and algorithms increase. Healthcare data is a good test bed for data mining. A great deal of data in health care is still being gathered and organized using pen and paper. In this research, we have used the BRHP database as the experimental study that consists of 25 attributes and 66,123 cases sampled (95,220 cases after SMOTE) from 236,549 cases.

In this research J48 decision tree algorithm is applied. The main aim of this research is to identify the best performing scenario of data mining the technique with knowing the most determining factor/attribute for the given dataset of the research.

Several models are developed as experimental analysis to outperform some of the J48 parameters. The models built allow us more flexibility with our output and can be more powerful weapons in our data mining arsenal. Most of the models developed is comparable and have closer performance accuracy for the issue under consideration. As we can see from experiment (scenario) 7 in figure 5.4, 90.30% predictive accuracy is obtained for the selected best model. That means 90.30% of the test data represents the majority class of the training set. The time required for computation and classification in this method is minimal.

In this research, the prediction rate of the J48 algorithm has revealed that mining the vital statistics data in BRHP is possible or applicable with 90% accuracy. To achieve for this performance accuracy use SMOTE that provides a new approach to the given data set. The results show that using the SMOTE approach can improve the accuracy of classifiers for a dataset.

Hence it is possible to conclude that the vital statistics data (death or mortality dataset) can be predicted by the application of classification technique (J48 decision tree algorithm) given the limitation of this study. In general, the result from this study is encouraging.

6.2 Recommendations

There are a number of different data mining algorithms that produce model to extract rules that can be used in decision support system. In addition to these algorithms, an application of SMOTE to data mining is important when class imbalance occurs in the target attribute to be predicted. There are also several topics to be considered further in this line of research.

A possibility for future work could be to implement a local database for the analysis where user can input data directly into their analysis, and based on the rule set, can deliver the answer back. That means classification can be done using all stored dataset in the database of the BRHP. This can be a handy tool for health practitioners (if possible to implement).

As this research work is an academic practical exercise, is should be considered as a preliminary effort to give insight into the application of data mining technology for the specified area of the research problem. Based on the result obtained, the researcher forwards the following recommendations for policy makers and future works.

- It is important to deploy the model built and selected with these techniques to use as a primary decision support in the identification of individual's mortality risk and grouping patterns.

- The concerned bodies for BRHP may need to investigate the applicability of data mining technology in more details and other details related comprehensive approach for strategic business needs.
- As stated earlier, the data set that is conducted for this research work is only a sample of eighteen years data with selected attributes. Hence it may be appropriate to consider the full dataset with all attribute as this gives broader opportunity to look into profound phenomena on it.
- In the researcher opinion, the number of experiments that are conducted in this research work is not ample enough to obtain best accuracy model using the BRHP database. Hence doing further researches shall be considered by using different algorithms of the same or different software with a number of scenarios.

REFERENCES

[1] Anthony, S. and Fauci, H. (ed) On Harrison's Principles of Internal Medicine. New York": McGraw-Hill, 1997

[2] Topics in Statistical Data Analysis Revealing Facts [Online] Available from: http://home.ubalt.edu/ntsbarsh/Bussiness-stat/home.html [Accessed 30 February 2011].

[3] Frawley, W. J. and Piatetsky, S. G. Knowledge Discovery in Databases, AAAI Press and MIT Press, Cambridge, MA. 1991

[4] Fayyad, U., Piatetsky, S. G. and Smyth, P. 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96),.AAAI Press, USA 1996

[5] Berhane, Y. and Byass, P. Butajira DSS Ethiopia , Department of Community Health, Faculty of Medicine Addis Ababa University and Department of Public Health and Clinical Medicine Umeå University, *INDEPTH Monograph*: Volume 1 Part C

[6] Xue Li and Osmar R. Z. Advanced Data Mining and Applications, Wuhan: Hubei Dictionaries Press. 2006.

[7] From Wikipedia, the free encyclopaedia [online] Available from: http://en.wikipedia.org/wiki/Data_mining#Methods [Accessed 23 February 2011].

[8] Michael, S., Taiki, S., Hua-Ching S., Charles, H., Steven, B. D., and Michael, J. O. Case Study: How to Apply Data Mining Techniques in a Healthcare Data Warehouse. 2007

[9] Arun, G. E. Application of Data mining in Medical Applications. Master's Thesis, Applied Science in Systems Design Engineering, the University of Waterloo, Waterloo, Ontario, Canada. 2004

[10] Ayalew, T. Electronic Vital Events Registration System (EVERS). Master's Thesis, FOI, Addis Ababa University, Addis Ababa. 2007

[11] Berhane, Y., Wall, S., Kebede, D., Emmelin, A., Enquselassie, F., Byass, P., Muhe, L., Anderson, T., Deyessa, N., Gossaye, Y., Hogberg, U., Alem, A., and Dahlblom, K. Establishing an epidemiological field laboratory in rural areas - potentials for public health research and interventions. *The Butajira Rural Health Programme 1987-1999. Ethiopian Journal of Health Development 13 Special Issue* (1-47). 1999

[12] Mark, T. Why Use Microsoft Data Mining? SQL Server 2008 Published. 2009

[13] Gossaye, Y., Deyessa, N., Berhane, Y., Ellsberg, M., Emmelin, M., Ashenafi, M., Alem, A., Negash, A., Kebede, D., Kullgren,G., and Hogberg U. Butajira Rural Health Program: Women's Health and Life Events Study in Rural Ethiopia. *Ethiopian Journal of Health Development* (17) *Special Issue*.

[14] Anagaw, S. Application of Data Mining technology to predict child mortality patterns:.The case of Butajira Rural Health Program (BRHP). Masters Thesis, FOI, Addis Ababa University, Addis Ababa. 2002

[15] Berry, J.A.M., and Linoff, G. *Data mining techniques-for marketing, sales and customer support*, New York: Wiley. 1997

[16] Lorase, T.D. Data Mining Methods and Models, Jhon Willy & Sons, USA 2006

[17] Nitesh, V. Chawla, Kevin, W. Bowyer, Lawrence, O. Hall, W. and Philip K. SMOTE: Synthetic Minority Over-sampling Technique. Department of Computer Science and Engineering, ENB 118. University of South Florida

[18] Huang, H., Tsai, W.T., Bhattacharya, S., Chen, X.P., Wang, Y., and Sun, J. Business rule extraction from legacy code, Proceedings of 20th International Conference on Computer Software and Applications, IEEE COMPSAC'96, 1996

[19] Wuthrich, B. Knowledge Discovery in Databases [Online] Technical Report CS-95-4, the Hong Kong University of Science & Technology, 1995. Available from: http://citeseer.nj.nec.com/89234.html [Accessed 30 February 2011].

[20] James R. R. Data-Mining-And-Importance-to-Achieve-Competitive-Edge-in-Business [Online]. Available from: http://ezinearticles.com/ [Accessed 5 March 2011]

[21] Han, J. and Kamber, M. Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers 2001.

[22] Data mining Research-DMR: [Online] A Promising Research Area. Available from: http://www.kdnuggets.com [Accessed 10 March 2011].

[23] Feldens, M. Towards a Methodology for the Discovery of Useful Knowledge – Combining Data Mining, Data Warehousing and Visualization', CNPq/Protem-cc Fase III (SIDI Project), Universidade Federal do Rio Grande do Sul, Brazil.CRISP-DM (2000) – 'CRISP-DM 1.0 – Step by Step data mining guide', CRISP-DM Consortium. 1998

[24] Collier, K., Sautter, D., Medidi, M., Morgan, J., Ratliff, M., Marjaniemi, C., Carey, B. and Abramo, P. A Perspective on Data Mining', Centre for Data Insight, Northern Arizona University, USA, 1998 pp 4-6.

[25] Lee S. W. and Kerschberg, L. A Methodology and Life Cycle Model for Data Mining and Knowledge Discovery in Precision Agriculture, Conference Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society Press, , San Diego, CA, USA 1998 pp. 2882-2887

[26] Ganesh, M., Kumar, V., Han, E., Shekhar, S.and Srivastava, J. 'Visual Data Mining: Framework and Algorithm Development', Department of Computing and Information Sciences, University of Minnesota, MN, USA 1996

[27] Kopanakis, I. and Theodoulidis, B. Visual Data Mining & Modeling Techniques, Centre of Research in Information Management (CRIM), Department of Computation, University of Manchester Institute of Science and Technology, UK 1999

[28] CRISP-DM– 'CRISP-DM 1.0 – Step by Step data mining guide', CRISP-DM Consortium. 2000

[29] Frank, S. Mine Your Own Data with the JDM API: Exploring the Java Data Mining API. 2005.

[30] Mark, F. Hornick, Erik, M., and Sunil, V. Book excerpt: Java Data Mining concepts, JavaWorld.com, 2007

[31] Morgan, K. Java Data Mining: [Online] Strategy, Standard, and Practice. 2007. Available from http:// www.mkp.com [Accessed 15 March 2011]

[32] Simoudis, E. Discovering Data Mining, Upper Saddle River, Prentice Hall PTR, New Jersey, USA 1998

[33] Last, M. and Kandel, A. Automated Perceptions in Data Mining [Online]. 2002. Available from :http://www.csee.usf.edu/~mlast/papers/perc_f1.pdf [Accessed 16 March 2011]

[34] Breiman, L., Freidman, J.H., Olshen, R.A. and Stone, C. J. Classification and Regression Trees. Wadsworth International Group, Belmont, CA. 1984

[35] Garner, S.R. WEKA: The Waikato Environment for Knowledge Analysis Proc" New Zealand Computer Science Research Students Conference, University of Waikato, Hamilton, New Zealand. 1995

[36] Chih-Lin, C. Medical decision support systems based on machine learning methods. Master's Thesis, the University of Iowa 2009

[37] Fayyad, U., Piatetsky – Shapiro, G., and Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data 2008

[38] Franky, D. C. Data Mining and Statistics in a pharmaceutical environment. SPS (Europe), Mechelen, Belgium 2006

[39] Lovleen K. G. and Rajni M. The Lure of Statistics in Data Mining. Guru Nanak Dev University, India, BBK DAV College for Women, India 2009

[40] Knowledge Discovery and Data Mining: Concepts and Fundamental Aspects. Available from http://www.worldscibooks.com/etext/5686/5686_chap1.pdf [Accessed 16 March 2011]

[41] Rea, A. Data Mining: [Online] An introduction Student Notes. 2002.Available from:http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.htm l [Accessed 19 March 2011]

[42] Tom, M. Machine Learning, McGraw Hill, 1997

[43] Kan, S. H. Metrics and Models in Software Quality Engineering, Addison-Wesley,1995

[44] Fayyad,U., Pazzani, M.J., Smyth, P., and Piatetsky, S. "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

[45] Sunil V. Using Java Data Mining to Develop Advanced Analytics Applications [Oline]. Available from http://java.sys-con.com/node/49091 [Accessed 19 March 2011]

[46] Greg, R. and Ellen, J. Mining your data for health care Quality improvement.SAS Institute, Inc., Cary, NC. 2009

[47] Benjamin, G. An Empirical Study of Machine Learning Algorithms Applied to Modeling Player Behavior in a .First Person Shooter. Video Game. Masters Thesis, University of Wisconsin – Madison. 2002

[48] Kim, H. and Loh, W. Y. Classification trees with unbiased multiway splits, Journal of the American Statistical Association, (96), pp. 589-604 2001

[49] Wilson A., Thabane L. and Holbrook A. Application of data mining techniques in pharmacovigilance. British Journal of Clinical Pharmacology. 2 (57), 127-134. 2003

[50] Raghavan, V., Deogun B., Jitender S. and Sover M. Data Mining: [Online] Trends and Issues. 2002. Available from: http://citeseer.nj.nec.com/138316.html [Accessed 30 March 2011].

[51] Shortliffe, E. H., Perrault, L. E. (ed) On Medical informatics: Computer applications in health care and biomedicine (2nd Edition). New York: Springer.2000

[52] Weiss, S. M., and Kulikowski, C. A. Computer Systems That Learn, Morgan Kaufmann Publishers, San Mateo. 1991

[53] Raymond, T. N. and Jian, P. B. Introduction to the Special Issue on Data Mining for Health Informatics. Department of Computer Science School of Computing Science Simon Fraser University and the University of British Columbia Vancouver, BC, Canada. 2008

[54] Tufte, E. Visual Explanations. Images and Quantities, Evidence and Narrative. Connecticut: Graphics Press. 1997

[55] Audain, C. Florence Nightingale [Online]. 2007. Available from: http://www.scottlan.edu/lriddle/women/ nitegale.htm [Accessed 30 March 2011]

[56] Benzler, J.; Herbst, A.J. and MacLeod, B. A reference data model for demographic surveillance systems [Online] INDEPTH. 1999. Available from http://www.indepth-network.org [Accessed 30 March 2011]

[57] Kanittha, V. Applying Knowledge Discovery in Databases in Public Health Data Set: Challenges and Concerns. University of Wisconsin-Madison School of Nursing. 2005

[58] Cheng, T.H., Wei, C.P. and Tseng, V.S. Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06) 2006

[59] Shillabeer, A. and Roddick, J. Establishing a Lineage for Medical Knowledge Discovery. ACM International Conference Proceeding Series. (311) 70, pp 29-37. 2007

[60] Health Grades, Inc. The Fourth Annual Health Grades Patient Safety in American Hospitals Study. 2007.

[61] Lavrac, N., Flach, P. and Zupan, B. Rule Evaluation Measures: A Unifying View, In Proceedings of ILP '99, LNAI 1634, 1999

[62] Hsu, W., Lee, M., Liu, B., and Ling, T. Exploration data mining in diabetic patient database. ACM. 2000 pp 430–436

[63] Wong, W.K., Moore, A., Cooper, G. and Wagner, M. What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Learning Research. (6) 2005. pp 1961-1998.

[64] Dibaba, A. Application of Data Mining technique to predict Household Health Seeking patterns: The case of Butajira Rural Health Program (BRHP). Masters Thesis, FOI, Addis Ababa University, Addis Ababa 2010

[65] Zhang D., Luan, Q. H. and Meiliu L. Mining California Vital Statistics Data. Department of Computer Science California State University, Clifornia 2008

[66] Andreeva, P., Dimitroval, M. and Radeva, P. Data Mining Learning Models and Algorithms For Medical Applications. Institute of Control and System Research,

Bulgarian Academy of Sciences. And Computer Vision Center, Autonomous University Barcelona, Barcelona, Spain 2005

[67] Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C. (ed) On Knowledge Discovery in Databases: An Overview. In Knowledge Discovery In Databases, AAAI Press/MIT Press, Cambridge, MA., 1991

[68] Shamebo, D. The Butajira Rural Health Project in Ethiopia: Epidemiological Surveillance for Research and Intervention in Primary Health Care..The Ethiopian Journal of Health Development, (8) Special Issue 1994

[69] Witten, T.H and Frank, E. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco 2000

[70] Habte, D. Population and health in developing countries Published by the International Development Research Centre (1) 2001

APPENDICES

Annex A: Sample BRHP dataset prepared for model building

No.	1: PA Nominal	2: ENVIR Nominal	3: REL Nominal	4: SEX Nominal	5: MARITAL Nominal	6: AGE Nominal	7: RSTART Nominal	8: YSTART Nominal	9: YEND Nominal	10: TIMEX Numeric	11: RELIG Nominal	12: LITER Nominal	13: EDUCAT Nominal
1	Mmeskan	H	HE	M	UK	'(42-5...	IN	'(1993.8-...	'(1993....	365.0	MU	IL	UK
2	Mmeskan	H	HE	M	MO	'(42-5...	MU	'(1998.9-...	'(2002....	1848.0	MU	LI	NO
3	Mmeskan	H	CH	M	TY	'(10.5-...	IN	'(1990.4-...	'(1993....	1095.0	MU	TY	NO
4	Mmeskan	H	CH	M	NM	'(10.5-...	XX	'(1995.5-...	'(1998....	1438.0	MU	LI	NO
5	Mmeskan	H	CH	M	NM	'(10.5-...	MU	'(1998.9-...	'(2002....	1145.0	MU	LI	NO
6	Mmeskan	H	CH	F	TY	'(10.5-...	BI	'(1988.7-...	'(1993....	2302.0	MU	TY	NO
7	Mmeskan	H	RE	F	NM	'(10.5-...	MU	'(1998.9-...	'(2000....	981.0	MU	UK	NO
8	Mmeskan	H	CH	F	TY	'(-inf-1...	XX	'(1995.5-...	'(1998....	1438.0	MU	TY	NO
9	Mmeskan	H	CH	F	TY	'(-inf-1...	MU	'(1998.9-...	'(2002....	1848.0	MU	UK	NO
10	Mmeskan	H	CH	F	TY	'(-inf-1...	MU	'(1998.9-...	'(1998....	249.0	MU	TY	NO
11	Mmeskan	H	CH	F	TY	'(-inf-1...	MU	'(1998.9-...	'(2002....	1783.0	MU	TY	NO
12	Mmeskan	H	HE	F	UK	'(73.5-...	IN	'(1990.4-...	'(1993....	1461.0	MU	IL	UK
13	Mmeskan	H	CH	M	UK	'(31.5-...	ST	'(-inf-198...	'(1993....	3045.0	MU	LI	UK
14	Mmeskan	H	SP	F	MO	'(31.5-...	CM	'(1998.9-...	'(2002....	1644.0	MU	LI	UK
15	Mmeskan	H	CH	F	UK	'(10.5-...	CM	'(1998.9-...	'(2002....	1644.0	MU	UK	UK
16	Mmeskan	H	CH	F	UK	'(10.5-...	CM	'(1998.9-...	'(2002....	1644.0	MU	UK	UK
17	Mmeskan	H	CH	F	TY	'(-inf-1...	BI	'(1995.5-...	'(2002....	3035.0	MU	UK	NO
18	Mmeskan	H	HE	M	MO	'(21-3...	MO	'(2000.6-...	'(2000....	51.0	MU	UK	NO
19	Mmeskan	H	CH	M	NM	'(21-3...	CM	'(1995.5-...	'(1998....	706.0	MU	LI	NO
20	Mmeskan	H	HE	M	MO	'(42-5...	XX	'(1995.5-...	'(1998....	1411.0	MU	LI	NO
21	Mmeskan	H	CH	M	NM	'(21-3...	MU	'(1998.9-...	'(2002....	1848.0	MU	LI	NO
22	Mmeskan	H	CH	M	NM	'(10.5-...	XX	'(1995.5-...	'(1998....	1438.0	MU	LI	UK
23	Mmeskan	H	CH	M	NM	'(10.5-...	MU	'(1998.9-...	'(2002....	1848.0	MU	LI	UK
24	Mmeskan	H	CH	M	TY	'(10.5-...	BI	'(1990.4-...	'(1993....	1572.0	MU	TY	NO
25	Mmeskan	H	CH	M	NM	'(10.5-...	MU	'(1998.9-...	'(2002....	1848.0	MU	LI	UK
26	Mmeskan	H	CH	M	TY	'(-inf-1...	BI	'(1998.9-...	'(1998....	70.0	MU	TY	NO
27	Mmeskan	H	HE	M	MO	'(63-7...	XX	'(1995.5-...	'(1998....	1411.0	MU	LI	UK
28	Mmeskan	H	SP	M	UK	'(52.5-...	ST	'(-inf-198...	'(1990....	1643.0	MU	LI	UK
29	Mmeskan	H	CH	F	UK	'(21-3...	ST	'(-inf-198...	'(1990....	1643.0	MU	TY	UK
30	Mmeskan	H	CH	M	UK	'(21-3...	LI	'(1990.4-...	'(1993....	1642.0	MU	LI	UK
31	Mmeskan	H	HE	M	MO	'(21-3...	MO	'(2002.3-...	'(2002....	84.0	MU	UK	NO
32	Mmeskan	H	CH	F	TY	'(10.5-...	CM	'(1995.5-...	'(1998....	706.0	MU	TY	NO
33	Mmeskan	H	CH	M	MO	'(31.5-...	XX	'(1995.5-...	'(1997....	989.0	MU	LI	UK
34	Mmeskan	H	CH	M	MO	'(31.5-...	MO	'(1997.2-...	'(2002....	2297.0	MU	LI	UK
35	Mmeskan	H	SP	F	UK	'(31.5-...	ST	'(-inf-198...	'(1990....	1643.0	MU	RE	UK
36	Mmeskan	H	RE	M	TY	'(-inf-1...	MU	'(1998.9-...	'(2000....	725.0	MU	LI	NO
37	Mmeskan	H	GP	F	WI	'(63-7...	CM	'(1995.5-...	'(1998....	399.0	MU	LI	UK
38	Mmeskan	H	CH	M	NM	'(10.5-...	IN	'(2002.3-...	'(2002....	721.0	MU	LI	NO

14: SOURCEW Nominal	15: ROOF Nominal	16: WINDOWS Nominal	17: RADIUS Numeric	18: ROOMS Numeric	19: HOUSEOWN Nominal	20: OXEN Nominal	21: TIMAD Numeric	22: LAT1 Numeric	23: LONG2 Numeric	24: DISTHOSP Numeric	25: STATUS Nominal
RI	TH	NO	4.0	1.651718	OW	UK	2.84737	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.651718	OW	UK	2.84737	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.651718	OW	UK	2.84737	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	4.0	1.0	OW	YE	4.0	8.09588	38.37672	3.5	Alive
RI	TH	NO	3.0	1.0	OW	UK	4.0	8.0979	38.37525	3.3	Alive
WU	TH	NO	4.0	1.0	OW	UK	5.0	8.11142	38.37478	1.8	Alive
WU	UK	YE	3.386331	1.0	OW	NO	3.0	8.10277	38.37312	2.8	Alive
WU	UK	YE	3.386331	1.0	OW	NO	3.0	8.10277	38.37312	2.8	Alive
WU	UK	YE	3.386331	1.0	OW	NO	3.0	8.10277	38.37312	2.8	Alive
WU	UK	YE	3.386331	1.0	OW	NO	3.0	8.10277	38.37312	2.8	Alive
WU	UK	NO	3.386331	1.0	OW	NO	2.0	8.10268	38.3731	2.8	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
WU	TH	NO	3.0	1.0	OW	UK	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	3.0	8.09963	38.3708	3.1	Alive
RI	TH	NO	3.0	1.0	OW	YE	2.0	8.09357	38.37648	3.8	Alive
RI	TH	NO	3.0	1.0	OW	UK	2.0	8.09357	38.37648	3.8	Alive
RI	TH	NO	3.0	1.0	OW	UK	2.0	8.09357	38.37648	3.8	Alive
RI	TH	NO	3.0	1.0	OW	UK	2.0	8.09357	38.37648	3.8	Alive
RI	TH	NO	3.0	1.0	OW	YE	4.0	8.09063	38.37417	4.1	Alive
PI	UK	NO	3.0	1.0	OW	NO	3.0	8.0982	38.37417	3.3	Alive
WU	TH	NO	2.0	1.0	OW	NO	3.0	8.0982	38.37417	3.3	Alive
WU	TH	NO	2.0	1.0	OW	NO	6.0	8.10248	38.37393	2.8	Alive
WU	TH	NO	3.386331	1.0	OW	NO	3.0	8.0982	38.37417	3.3	Alive
RI	TH	NO	2.0	1.0	OW	UK	6.0	8.10248	38.37393	2.8	Alive
WU	TH	NO	3.386331	1.0	OW	NO	3.0	8.0982	38.37417	3.3	Alive
WU	TH	NO	2.0	1.0	OW	NO	3.0	8.0982	38.37417	3.3	Alive
WU	TH	NO	2.0	1.0	OW	NO					

Annex B: A partial decision tree generated for BRHP dataset

```
J48 pruned tree
------------------

ROOF = TH
|  TIMEX <= 1380.978386
|  |  RSTART = IN: Alive (4001.0/402.0)
|  |  RSTART = MU
|  |  |  YEND = '(-inf-1988.7]': Dead (132.0)
|  |  |  YEND = '(1988.7-1990.4]': Dead (311.0)
|  |  |  YEND = '(1990.4-1992.1]': Dead (555.0)
|  |  |  YEND = '(1992.1-1993.8]': Dead (730.0)
|  |  |  YEND = '(1993.8-1995.5]': Dead (1134.0)
|  |  |  YEND = '(1995.5-1997.2]': Dead (2401.0/35.0)
|  |  |  YEND = '(1997.2-1998.9]': Dead (3754.0/65.0)
|  |  |  YEND = '(1998.9-2000.6]'
|  |  RSTART = XX
|  |  |  RELIG = MU
|  |  |  |  |  MARITAL = UK: Dead (61.0/25.0)
|  |  |  |  |  MARITAL = MO
|  |  |  |  |  |  EDUCAT = UK: Dead (1276.0/122.0)
|  |  |  |  |  |  EDUCAT = NO
|  |  |  |  |  |  EDUCAT = PR: Alive (15.0/3.0)
|  |  |  |  |  |  EDUCAT = SE: Alive (2.0)
|  |  |  |  |  MARITAL = TY: Dead (2100.0/239.0)
|  |  |  |  |  MARITAL = NM
|  |  |  |  |  |  EDUCAT = UK
|  |  |  |  |  |  |  |  ENVIR = H: Dead (257.0/94.0)
|  |  |  |  |  |  |  |  ENVIR = U: Alive (1.0)
|  |  |  |  |  |  |  |  ENVIR = L: Alive (153.0/59.0)
|  |  |  |  |  |  EDUCAT = NO: Alive (161.0/30.0)
|  |  |  |  |  |  EDUCAT = PR: Alive (16.0/3.0)
|  |  |  |  |  |  EDUCAT = SE: Alive (3.0)
|  |  |  |  |  MARITAL = WI: Dead (621.0/51.0)
|  |  |  |  |  MARITAL = PO: Dead (123.0/53.0)
|  |  |  |  |  MARITAL = DI: Dead (25.0/10.0)
|  |  |  |  |  MARITAL = SE: Dead (12.0/4.0)
|  |  |  RELIG = OC
|  |  |  |  AGE = '(-inf-10.5]': Dead (47.0/19.0)
|  |  |  |  AGE = '(10.5-21]': Alive (129.0/52.0)
|  |  |  |  AGE = '(21-31.5]': Alive (138.0/19.0)
|  |  |  |  AGE = '(31.5-42]': Alive (67.0/15.0)
|  |  |  |  AGE = '(42-52.5]': Alive (51.0/22.0)
|  |  |  |  AGE = '(52.5-63]': Dead (41.0/20.0)
|  |  |  |  AGE = '(63-73.5]': Dead (50.0/12.0)
```

```
| | | | AGE = '(73.5-84]': Dead (33.0/8.0)
| | | | AGE = '(84-inf)': Dead (28.0/2.0)
| | | RELIG = UK: Dead (84.0/33.0)
| | | RELIG = CH: Alive (91.0/25.0)
| | | RELIG = OT: Alive (1.0)
| | RSTART = ST
| | | AGE = '(-inf-10.5]': Alive (0.0)
| | | AGE = '(10.5-21]': Dead (299.0/70.0)
| | | AGE = '(21-31.5]'
| | | AGE = '(31.5-42]': Alive (384.0/83.0)
| | | AGE = '(42-52.5]': Alive (162.0/70.0)
| | | AGE = '(52.5-63]': Dead (130.0/55.0)
| | | AGE = '(63-73.5]': Dead (108.0/39.0)
| | | AGE = '(73.5-84]': Dead (107.0/27.0)
| | | AGE = '(84-inf)': Dead (94.0/15.0)
| | RSTART = MO
| | | YEND = '(-inf-1988.7]': Dead (3.0)
| | | YEND = '(1988.7-1990.4]': Dead (10.0)
| | | YEND = '(1990.4-1992.1]': Dead (12.0)
| | | YEND = '(1992.1-1993.8]': Dead (8.0)
| | | YEND = '(1993.8-1995.5]': Dead (14.0)
| | | YEND = '(1995.5-1997.2]': Dead (50.0/21.0)
| | | YEND = '(1997.2-1998.9]': Dead (47.0/18.0)
| | | YEND = '(1998.9-2000.6]': Alive (287.0/55.0)
| | | YEND = '(2000.6-2002.3]': Alive (88.0/14.0)
| | | YEND = '(2002.3-inf)': Alive (259.0/13.0)
| | RSTART = LI: Alive (426.0)
| | RSTART = RO: Dead (19.0/9.0)
| | RSTART = WA: Dead (0.0)
ROOF = CO
| TIMEX <= 1377
| | RSTART = IN: Alive (1637.0/55.0)
| | RSTART = MU: Alive (1557.0/258.0)
| | RSTART = XX
| | | AGE = '(-inf-10.5]': Alive (49.0/22.0)
| | | AGE = '(10.5-21]': Alive (156.0/37.0)
| | | AGE = '(21-31.5]': Alive (165.0/18.0)
| | | AGE = '(31.5-42]': Alive (82.0/14.0)
| | | AGE = '(42-52.5]': Dead (56.0/27.0)
| | | AGE = '(52.5-63]': Alive (25.0/9.0)
| | | AGE = '(63-73.5]': Dead (27.0/6.0)
| | | AGE = '(73.5-84]': Dead (20.0/5.0)
| | | AGE = '(84-inf)': Dead (12.0/2.0)
```

ADDIS ABABA UNIVERSITY
College of Health Sciences
School of Public Health
Ethical Clearance Form

Version 01.Nov.2010

Date: / 09 / 05 / 11 /
Ref.No. SPH/0194/03

Project number / 0060/

Date of approval (D/M/Y) /29 /11/10/	
Project Title: Mining vital statistics data the case of Butajira Rural Health Program	
Name of PI Tadesse Beyene	Phone Number
Institution	School of Public Health
Department	
Decision of Research and Ethics Committee:	✓ Approved ☐ Approved with Recommendation ☐ Resubmission ☐ Disapproved
Valid until	May 1, 2011 – Oct. 1, 2011

Dean, School of Public Health

Signature _____

Date / 9 / 5 /

A.A.U. College of Health Sciences Dean's Office School of Public Health