ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND

SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNOLOGY IN PREDICTING THE SEROPREVALENCE OF HBV,HCV,HIV; THE CASE OF THE NATIONAL BLOOD BANK OF ADDIS ABABA,ETHIOPIA

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

By

HAFTOM GEBREGZIABHER

July, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

APPLICATION OF DATA MINING TECHNOLOGY IN PREDICTING THE SEROPREVALENCE OF HBV,HCV,HIV; THE CASE OF THE NATIONAL BLOOD BANK OF ADDIS ABABA,ETHIOPIA

By

HAFTOM GEBREGZIABHER

July, 2011

Name and Signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	____	Chairperson _____	_____
Million Meshesha	(PhD)	Advisor(s), _____	_____
Wubegziaher Mekonen	(MA)	Advisor(s), _____	_____
Dereje Teferi		Examiner, _____	_____

DEDICATION

This paper is dedicated to my sister, W/ro Genet Gebregziabher.

DECLARATION

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

_____

Haftom Gebregziabher

July, 2011

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

Million Meshesha (PhD) _____

Wubegziabher Mekonen (MA) _____

ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Million Meshesha,and Ato Wubegziabher Mekonen for their devotion and constant support during my work. This research wouldn't have been materialized without their time wise and constructive comments.

My special thanks goes to Dr. Kifle Mekonen advisor at the Federal Ministry of Health for bringing the research area to my attention.

I would also want to extend my gratitude to staff members of the blood transfusion service and data management unit of the Ethiopian national blood bank service.

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

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ABBREVIATIONS

AAU: Addis Ababa University

AIDS: Acquired Immunious Deficiency Syndrome

CRISP-DM: Cross Industry Standard Process for Data Mining

DM: Data Mining

ENNBS: Ethiopian National Blood Bank Service

ERCS: Ethiopian Red Cross Society

FMOH: Federal Ministry of Health

HAPCO: HIV/AIDS Prevention and Control Office

HBV: Hepatitis B Virus

HIV: Human Immunious Virus

HCV: Hepatitis C Virus

KDD: Knowledge Discovery in Databases

KDP: Knowledge Discovery Process

OLAP: Online Analytical Processing

SMOTE: Synthetic Minority Over-sampling Technique

TTI's: Transfusion Transmittable Infections

ABSTRACT

Recent advancements in communication technologies, on the one hand, and computer hardware and database technologies, on the other hand, have made it easy for organizations to collect, store and manipulate massive amounts of data. As stated by Deogan, these large databases contain potential gold mine of valuable information, but it is beyond human ability to analyze substantial amounts of data and extract meaningful patterns. As the volume of data increases, the proportion of information in which people could understand decreases substantially. The applications of learning algorithms in knowledge discovery are promising and they are relevant area of research offering new possibilities and benefits in real-world applications such as blood bank data warehouse. The availability of optimal blood in blood banks is a critical and important aspect in a Blood transfusion service. Blood banks are typically based on a healthy person voluntarily donating blood used for transfusions. The ability to identify regular blood donors enables blood bank and voluntary organizations to plan systematically for organizing blood donation camps in an efficient manner.

The objective of this study is to explore the immense applicability of data mining technology in the Ethiopian National Blood Bank Service by developing a predictive model that could help in the donor recruitment strategies by identifying donors that are at risk of TTI's which can help in the collection of safe blood group which in turn assists in maintaining optimal blood.

The analysis has been carried out on 14575 blood donor's dataset that has at least one pathogen using the J48 decision tree and Naïve Bayes algorithm implemented in Weka. J48 decision tree algorithm with the overall model accuracy of 89 % has offered interesting rules.

From the total of 156729 consecutive blood donors, 14757 (9.41%) had serological evidence of infection with at least one pathogen and 29 (0.19%) had multiple infections. The overall seroprevalence of HIV, HBV and HCV was 2.29%, 5.23%, and 2.30% respectively .The seropositivity of TTI's was significant in business owners, students, civil servants, unemployed individuals, drivers and age groups 25 to 34and 35 to 44 years.

CHAPTER ONE

INTRODUCTION

1.1 Background

These days, we are witnessing the development of a new chapter in the information revolution caused by the junction of information and communication technologies. The new technology has radically changed our society and economy. In information storage and retrieval activities, technology has the potential to realize the ultimate dream of the information retrieval specialist: to make information available to any person, when and where it is required(1). As Bigus (2) stated, over the last four decades, the use of computer technology has evolved from gradual automation of certain business operations, such as accounting and billing, into today's integrated computing environments, which offer end-to-end automation of all major business processes. Not only the computer technology has changed, but also how that technology is viewed and how it is used in business has changed.

More than ever, in the health care sector, technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable(3). Although the capabilities to collect and store data in large computer databases has increased significantly, the relational database technology of today offers little functionality to process and explore data and establish a relationship or pattern among data elements that are hidden or previously unknown(4). Prather (3) noted that, although health care databases have accumulated large quantities of information about patients and their medical conditions, there are only few tools to evaluate and analyze this clinical data after it has been captured and stored. The author further stated that evaluation of stored clinical data might lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management.

As Fayyad and Piatetsky (5) wrote, the traditional method of turning data into knowledge relies on manual analysis and interpretation. The writers further stated that, for example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health-care organization; this report can be used for effective future decision making and planning for better health-care management.

Specifically, as data volumes grow dramatically, data analysis based on manual methods is becoming completely impractical in many domains. Prather (3)argued that to evaluate and analyze data stored in large databases, new techniques and methods are needed to search large quantities of data and to discover new patterns and relationships hidden in the data. It is due to these challenges of searching for knowledge in relational databases and our inability to interpret and digest these data as readily as they are accumulated, which has created a need for a new generation of tools and techniques for automated and intelligent database analysis. Consequently, the discipline of knowledge discovery in databases (KDD), which deals with the study of such tools and techniques, has evolved into an important active area of research(3).

According to Larvac (6), in the health care sector, the widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer assisted analysis. Such an extensive amount of data gathered in medical databases require specialized tools and methods that can be used to discover new information and knowledge which is useful in decision making and problem solving.

Faced with the tremendous economic and competitive pressures, the health-care industry has started to mine its data to minimize costs, enhance quality and save lives. In support of this notion, Bresnahan (7) argued that one way in which data mining is helping health-care providers cut costs and improve care is by showing which treatments have been most effective. For example, once hospital administrators recognize that stroke patients are less likely to develop respiratory infections if they can swallow properly, they can educate their physicians and

institute a standard policy to identify and provide therapy to those who have difficulty of swallowing.

The data mining process which serves as a means of searching previously unknown, actionable information from large databases can be used to improve the quality, efficiency and care of patients which is known in the health care industry as "outcomes measurements." An outcome measurement involves examining clinical encounter information, insurance claims and billing data to measure the results of past treatment and process(8). Bresnahan (7) further stated that since many of the issues and problems associated with outcomes measurement apply to data mining, health care providers can identify areas of improvement or capitalize on successful methods.

Blood donation and transfusion service is an indispensable part of contemporary medicine and healthcare(9). As a consequence, it is of vital importance to coordinate and administer various activities involved in blood donation and transfusion service. Nevertheless, blood management has been recognized as a challenging task: the life-threatening nature of blood products entails the punctilious administration while its perishable nature necessitates the timely processing. At the same time, the decentralized affairs involved in this procedure further complicate the effective administration of blood donation and transfusion service. Fortunately, such terrific challenge has been considerably alleviated with the development of information and computer technology. As a matter of fact, many successful and remarkable achievements have been reported in the field of computerized management of blood donation and transfusion service(10).

Nowadays the effective use of computer and information technology in blood bank system generally refers to acquiring, validating, storing, and circulating various data and information electronically in blood donation and transfusion service. Given the top priority of concerns on blood transfusion security, most reported systems are particularly devoted to the issues of data credibility, information consistency, and system reliability, etc. Even official implementation and evaluation guidelines pay little attention on the topics other than security and reliability in blood bank information systems(11).However, from the perspective of blood bank staffs, they often seek more support from the blood bank information systems other than inputting and retrieving

historical data only. At least, such system should be able to assemble the heterogeneous data into legible reports for appropriate decision making support.

It is known that any reasonable decision should comply with the objective data and subject to the supervision of knowledge. From effective donor screening to optimal blood dissemination, those electronic data in a blood bank information system indeed can contribute to various blood bank decisions. Thus, for a competent blood bank information system, it is not a trivial task to develop the effective decision support modules. Data mining and artificial intelligence among others are the tools that are providing an efficient way of data analysis for better use of data collected in blood banks(11).

As Butch (12) pointed out contemporary blood bank information systems are more interested in data analysis so as to predicatively optimize blood bank resource allocation and management. In the course of data analysis process computerized decision making support are built on either the practical experience from blood bank professionals or the implicit knowledge by data mining and knowledge discovery.

Santhanam and Shyam (10) argued blood banks (in the developing countries context) are typically based on a healthy person voluntarily donating blood and are used for transfusions or made into medications. The authors added the use of data mining technology in blood donation behavior has promising results in identifying regular blood donors which enables blood banks and voluntary organizations to plan systematically for organizing blood donation camps in an effective manner.

Rea (13) wrote data mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data corpus such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the objects in the database if the database is a faithful mirror of the real world.

Data mining includes major tasks such as classification, clustering and association rule discovery. In classification task, data can be defined in terms of attributes, one of which is the class. It finds a model for class attribute as a function of the values of other (predictor) attributes, such that previously unseen records can be assigned a class as accurately as possible. Decision tree, neural networks and bayes are the common method used for classification task.

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. It is mapping a data item into one of several clusters which are not pre-specified but are determined from the data. Clusters are formed by finding natural groupings of data items based on similarity matrices, proximity considerations and probability measures(14). There are partitioning algorithms (such as k-means and k-medoid clustering) and hierarchical algorithms (such as divisive and agglomerative clustering to find natural groupings)

In large databases, there are so many patterns that the user can never practically think of the right questions to ask although rich and interesting patterns that can be expressed and discovered are available. The task of data mining which interrogates the databases in search of such interesting rules is called association rule discovery(15).Using apriori and FP-Growth association rule discovery discovers patterns that can show the frequency of occurance of data items.

The application of data mining, in medical and health data is challenging and intriguing (Abidi & Goh, 1998; Brossette et al., 1998; Cios & Moore, 2002). The datasets usually are very large, complex, heterogeneous, and hierarchical and vary in quality. Data preprocessing and transformation are required even before mining and discovery can be applied. Sometimes the characteristics of the data may not be optimal for mining. The challenge here is to convert the data into appropriate form before any mining can begin. Furthermore, a system which is quick and correct on some small training sets, could behave completely different when applied to a larger database. A data mining system may work perfect for consistent data and perform significant worse when a little noise is added to the training set.

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 in its early inception when it comes to the Ethiopian health care sector. Given experiences elsewhere in terms of the benefits acquired in applying data mining technology in the health care sector, it is only proper to explore the relevance and potential advantage of such a state of the art technology in the Ethiopian health-care context.

1.2 Statement of the Problem and Justification

The two crucial issues related to blood transfusion in the developing world, particularly Africa, are blood shortages and unsafe blood(16), which all too frequently lead to serious health consequences such as death from postpartum hemorrhage or the transmission of life-threatening infections such as HIV and Hepatitis. These ill health consequences could be preventable through actions to improve blood safety and availability.

Blood services in Ethiopia have for the past 30 years been mainly provided by the Ethiopian Red Cross Society (ERCS) through its 12 regional blood banks with replacement and directed donations in 35% of its 126 hospitals countrywide. However, there has been inadequacy and in-equitability in access to safe blood by the population, particularly in the regions. Only 24,000 units of blood were collected in 2004 (i.e. 0.3 units/1000 people) and of these 17,000 units (71% of the total) were collected from Addis Ababa. The shortage of blood supplies were more evident for the vast majority of the population (about 96%) residing outside Addis Ababa.

On top of the shortage of blood donated nationwide, the prevalence of the major TTI's at the national blood bank(Hepatitis B Virus=5.23%,HIV=2.29% and Hepatitis C Virus=2.30%) indicates there is a high prevalence. This magnitude dictates a strict transfusion transmission service should be made in the course of blood donation process. Cognizant to the fact that the majority of the population (96%) is residing outside Addis Ababa, and the coverage of the

national blood bank service is limited to the major cities (such as Mekelle,Bahridar,Assosa) of the nation and this results the collection of safe and optimal blood inadequate.

Above all, the whole blood transfusion process demands a well trained health professionals and different laboratory kits to examine and detect the life-threatening infectious diseases. However, in countries like Ethiopia where the health care service is being challenged with adequate resources; the public health services such as the blood bank service remains in its low coverage. Hence due to the resource constraints the collected blood is partly unsafe and always below the demand(17).

Not less notable some of the collected bloods are discarded if samples are found to have at least one TTI's which adds its toll on the blood shortage. This happens because there is no other mechanism other than the physical screening to identify certain blood group patterns as susceptible to one or more TTI's. Given the demographic characteristics of the blood donors data; there is a need to explore the possibility of predicting future trends and outcomes of the blood donor which can possibly minimize the blood being discarded and know the safest blood group which helps the blood donation advocacy to be geared towards donors who are safe from infections.

Research conducted by Santhanam and shyam (10) on a blood bank dataset on application of CART algorithm in blood donors classification argued blood banks in the developing world context are typically based on healthy person voluntarily donating blood and is used for transfusions. The ability to identify regular blood donors enables blood banks and voluntary organizations to plan systematically for organizing blood donation camps in effective manner.
Other researches' have been done in the Ethiopian Blood Bank such as Seroprevalence of HIV, HBV, HCV and syphilis infections among blood donors at Gondar University Teaching Hospital, Northwest Ethiopia(18): it is observed that a declining trends over a period of five years and similar research to predict the prevalence of HBV, HCV and malaria parasites among blood donors in Amhara and Tigray regional states (19) but because of the time gap and the prediction capacity of techniques such as data mining technology further researches to generate novel

knowledge are evident. The data warehouse available at the national blood bank is used for keeping the records of donors' data and reporting statistical description of donors' donation proportion annually by age, sex, occupation and etc,. However, due to the limitation of statistical analysis such as their limited ability to learn new knowledge from existing data and their primary focus to test a given hypothesis, mining the pattern of blood and infectious diseases data is an opportunity to explore hidden knowledge and inform planning of health programmers to guide advocacy efforts. Given the demographic characteristics of donors' data mining can uncover important data patterns, contributing greatly to business strategies in providing a novel knowledge that can be used as a base for guidance and decision making. It is therefore the aim of this study to assess the potential applicability of data mining technology to predict the blood donation patterns that were identified with at least one known TTI's and predict the more safest group so that blood can be mobilized from group that has less risk of TTI's.

In the course of the research work this study is intended to give answer to the following research questions.

To what extent data mining helps in discovering patterns and knowledge for predicting the seroprevalence of TTI's.

What data mining algorithms and models are more suitable for predicting patterns among attributes of demographic characteristics of donors?

Which age group, locations , blood donation type are the most susceptible to at least one TTI's and exhibit similar trends for certain diseases?

1.3 Objectives

1.3.1 General objective

The main objective of this study is predicting the seroprevalence, and risk factors of HIV, HBV, and HCV infections among blood donors at the National Blood Bank of Ethiopia using data mining technology to extract useful information about the blood donors' characteristics and generate a new knowledge and patterns that help collection of safe blood.

1.3.2 Specific objectives

In order to achieve the general objective, the following specific objectives were attempted in the present research:

Assess different classification, clustering and association rules mining application algorithms.

Select and extract the data set required for analysis from the Ethiopian National Blood Bank.

Preprocess: preprocessing data in order to have a cleaned dataset that is suitable for any data mining algorithm.

Train and build data mining models that help for predicting the seroprevalence of HIV, HBV, HCV and syphilis at the blood bank.

Evaluate the performance of the data mining model using test data set and report findings.

1.4 Scope and limitation of the Research

This research is conducted based on the data obtained from the National Blood Bank Service of Ethiopia centered at Addis Ababa. Considering that the national blood bank is a representative of the other sub branches of the blood bank, a retrospective analysis of consecutive blood donors' records covering the period between September 2002 and December 2010 was used for analysis. The proposed study was intended to mine data that holds records of communicable diseases particularly Hepatitis B Virus, Hepatitis C Virus, HIV and Syphilis but because of medical reasons that Syphilis virus becomes inactive after three days following the actual donation process and it is not recorded in the database. Hence records having the HBV, HCV and HIV were used for analysis purpose.

Predictive modeling data mining task is performed for the extraction of unknown knowledge and interesting patterns that can be used as a base in setting rules in the donor recruitment process that can enhance the collection of safe blood.

1.5 Significance of the Study

The discovery of transfusion-transmissible infections (TTI's) has heralded a new era in blood transfusion practice worldwide with emphasis on two fundamental objectives, safety and protection of human life. Blood safety remains an issue of major concern in transfusion medicine in Ethiopia where national blood transfusion services and policies, appropriate infrastructure, trained personnel and financial resources are inadequate(18).

Continuous monitoring of the magnitude of transfusion-transmissible infections in blood donors is important for estimating the risk of transfusion and optimizing donor recruitment strategies to minimize infectious diseases transmission(18).

Outcomes measurement using data mining helps health-care institutions to evaluate their doctors and facilities. Physicians and hospitals benefit from knowing how they compare with their peers, and the parent company saves money by getting all of its employees up to par. Along the same lines, outcomes measurement by using data mining technology also lets caregivers identify people statistically at risk for certain ailments so that they can be treated before the condition escalates into somewhat expensive and potentially fatal(1).
Data mining and knowledge discovery can close the loop between clinical data mining capture and evidence-based decision support by facilitating the conversion of clinical data into evidence for future decision(20).

The Ethiopian National Blood Bank has a data warehouse that keeps track of the daily blood collected and distributed to the different hospitals across the nation. Mining the pattern of blood donated data will be an opportunity to explore hidden knowledge and inform health programmers to guide advocacy efforts and interventions at the national level. Given the data with different demographic characteristics mining the donor's data would provide information on which age group, sex, geographic location and type of blood donation are less vulnerable to infectious diseases. This may help the blood bank to design a targeted strategy to consult the group which is more exposed to infection and collect the safest blood which is free of any transfusion transmittable infectious diseases. Thus it paves a way for government intervention and directions to be adjusted based on available information. Furthermore, results of the data

mining will initiate further researches to be undertaken that were not done due to the limitation of the data on hand.

1.6 Methodology

The word methodology refers to a documented approach which is used to perform activities in a manner which is coherent, consistent, accountable and repeatable. Methodology is a process that mainly consists of intellectual activities. Usually only the end goal of the methodological process is manifested as the product or result of the physical work(21).

It is evident that certain set of steps are usually required to accomplish a certain task. These set of steps could guide which activity to do first and keep on doing in a chronological order. The choice of following the set of steps depends on how one is familiar with them and depending on the immense benefits they offer compared with others

1.6.1 Study design

Data mining problems like any other problem domains need to follow series of steps which are clear, consistent, repeatable and understandable usually termed as process models. The concept of a process model is used to formalize the knowledge discovery processes (KDPs) within a common framework. The model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the project. This in turn results in cost and time savings, better understanding, and acceptance results of the task. We need to understand that such processes are nontrivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners(22).

Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa(22).

Because of the fact that, this research is an academic research and results of the research can possibly be deployed to bring a solution, a hybrid knowledge discovery process model (KDD) which takes lessons both from the industrial and academic models is being used. It was

developed based on the CRISP-DM model by adopting it to academic research. replacing the modeling step with data mining process(23).

Figure 1.1 shows descriptions of the six steps of the KDD process model(22).

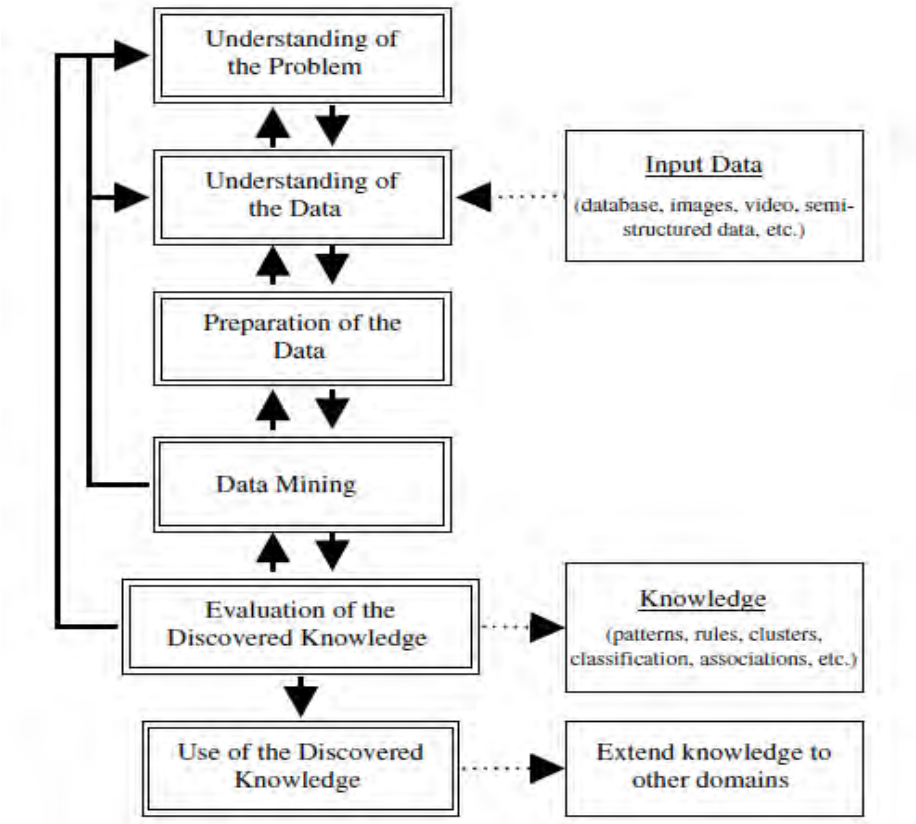

Figure 1.1 the KDD process model.

Understanding the problem domain

This initial phase focuses on understanding the project objectives and requirements from a business perspective which involves working closely with domain experts to define the problem and determine the project goals, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. Basically in order to have an insight in the overall business objectives and workflows of the Ethiopian National Blood Bank an intensive interview with dedicated staff members, document reviews and observations were made(23).

Understanding of the data

This step includes collecting sample data and deciding which data, including format and size, are needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals(23).

Once the business problem and the objective are clearly defined, the next step is to take time to understand what kind of data are available and which data is the most appropriate and has the most predictive power that can address the objective. Usually the data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. In order to understand the nature of the data and the attributes at the ENBBS database brief explanation was made by the database administrator. The information about each donor are stored in a database under different trees such as table “donor”, table ”infectious” and table ”Discard”. The record of the table “infectious” which holds the information about each donor with at least one of the TTI’s and a sample data that represents the safe group were used for analysis. Since Data mining algorithm need an ideal substantial data 16864 records were taken for analysis.

Preparation of the data

This step concerns deciding which data are used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data is further processed by feature selection. In fact the ideal practice for variable selection is to take all the variables in the database, feed them to the data mining tool and let it find those which are the best predictors. However, blindly including extraneous columns can lead to incorrect models although in principle some data mining algorithms can automatically ignore irrelevant variables and properly account for related

(covariant) columns. Therefore after consulting the database administrator of the ENBBS about the meaning of the attributes added by data analysts for easy data analysis purpose, the researcher came to learn the variables used were all important. However, because of the variation in woreda and kebelle attributes, they were merged in to subcity attribute. This reduces the number of attributes from 16 to 14. Missing values used for this research were handled by the global constant assigning the missing values the same constant as the missing values were few.

Data mining

It is only at this point that one invokes data mining models and tools to interrogate the data and convert it into knowledge for decision making. Two Crows Corporation(23), emphasizes the point that data mining model building is an iterative process. At this stage, we select a particular data mining method that matches the goals of the data mining process defined in the first step. However, the details of building and training a model vary from technique to technique and hence there are no blue print procedures. For this reason prior to training and building a model, a classification data mining technique such as j48 decision tree, bayes has been used in Weka 3.72 open source software for prediction purposes so as to give answer to the predefined objectives

Evaluation of the discovered knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared. The fact that, the classification capability of any model depends on the learning ability of the data sets provided, the model with the highest true positives and less false negatives means there is a likelihood of classifying the data sets correctly making the precision and the recall values to be maximum(24). Hence, the model with a maximum precision and recall value was chosen as an evaluation parameter. So in this research J48 algorithm was chosen having a better prediction and accuracy.

Use of the discovered knowledge

Creation of the knowledge is not the end of the study. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained always needs to be organized and presented in a way the customer can use it(24).

1.6.2 Study variables

The following variables are the attribute values used in the ENBBS database to keep track of the donors' record and used as study variables for this research.

- Age,,Sex,Occupation,Rh,Region,Subcity,Weight,Blood type(ABO),Dateofdonation,Type of donation,Site,HBV,HCV,HIV.

1.6.3 Ethical clearance

The study was carried out after getting clearance from the ethical clearance committee of the School of Public Health in Addis Ababa University. Data was collected after getting permission from the National Blood Bank of Ethiopia. To increase confidentiality the data was made be anonymous.

1.6.4 Quality of data management

Data quality is the reliability and effectiveness of data, particularly in a data warehouse. Before using data files exported from database to excel from as they were, cross checking the data files with their corresponding original data files in database were done. The exported files were cross checked again with the individual queries and in the course of cross checking the data files, it has been found that each pair of interrelated files contained identical data.

1.6.5 Dissemination of the research findings

Result of the research will be communicated through annual students and staff research conference in the University of Addis Ababa, national conference of Ethiopian Public Health Association,Ethiopian National Blood Bank Service and will be sent to journals based on their publishing standards and rules.

1.7 Organization of the Paper

This paper is divided 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, and methodologies adopted for the study.

The second chapter mainly revolves around the technology to be applied on this research project. Literature is reviewed on the different data mining models, techniques used, and the application of data mining technology in the health care sector.

The third chapter focuses on the business and data understanding of the ENBBS. This chapter includes the activities undertaken to understand the business functionality of the organization and the data collection, data preprocessing and statistical summarizations of attributes under study.

The fourth chapter is concerned with the detail techniques and algorithms followed in the knowledge discovery process.

The fifth chapter provides discussions about the different data mining steps that were undertaken in this research work. This includes model selection, building and evaluating and interpreting results obtained from j48.
The last chapter is devoted for the final conclusions and recommendations based on the research findings.

CHAPTER TWo

REVIEW OF LITERATURE

During World War II and the immediate post war period the demand for blood and blood components in the USA increased substantially(19, 25). This resulted in the establishment and growth of blood banks transfusion services and other blood laboratory support services. Despite its common occurrence, blood transfusion often raises concerns among patients and clinicians about potential infections.

Blood as defined in the medical dictionaries is a fluid that circulates through the heart, arteries, capillaries, and veins and is the chief means of transport within the body. It transports oxygen from the lungs to the body tissues, and carbon dioxide from the tissues to the lungs. It transports nutritive substances and metabolites to the tissues and removes waste products to the kidneys and other organs of excretion. It has an essential role in the maintenance of fluid balance(26). Blood transfusion is the transfer of blood or blood components from one person (the donor) into the bloodstream of another person (the recipient).

Ethiopian Red Cross Society (ERCS) was established on the eve of Ethio-Italian war of 1936-40, by government decree of July 8, 1935 and recognized by the International Red Cross and Red Crescent (IRCRC) the same year; it joined the International Federation in 1950. In October 1947 , the society was granted its first charter which was issued in the "Negarit Gazeta"(27).

The Ethiopian Red Cross and Red Crescent Society as member of the international federation firmly respects the fundamental principle of humanity, impartiality, neutrality, independence, voluntary service, unity and universality (28).
It is envisioned to inspire, encourage, facilitate and promote at all times all forms of humanitarian activities by National Societies, with a view to preventing and alleviating human suffering, and thereby contributing to the maintenance and promotion of human dignity and peace in the world.

The ERCS activities as part of the International Federation's programmes are grouped into four main core areas: promoting humanitarian principles and values; disaster response; disaster preparedness; and health and care in the community(28).

Too many people die as a result of no access to even the most basic health services and elementary health education. Health and community care has become a cornerstone of humanitarian assistance, and accounts for a large part of Red Cross Red Crescent spending. Through these programmes, the Federation aims to enable communities to reduce their vulnerability to disease, and prepare for and respond to public health crises (28).

The Ethiopian Red Cross society National Blood Bank Services (ERCS-NBBS) as one of the core activities of the ERCS is the sole organization providing blood bank services across the country since its establishment in 1969,with its central blood bank located at Addis Ababa, and eleven regional blood banks found in Adama, Harar, Diredawa, Jijiga, Yirgalem, Arbaminch, Jimma, BahirDar, Gondar,dessie and Mekelle(17).

Though the blood and blood products demand of the country in the year 2003 is estimated to be 75,000 to 80,000 units per year, the Ethiopian National Blood Bank cannot meet this demand at present. Hence, the ERCS-NBBS in collaboration with other major partners such as Federal Ministry of Health, World Health Organization and Center for Disease Control (FMOH, WHO, CDC) is now actively working so as to meet this demand. Besides to this the establishment of new blood banks and the promotion of voluntary blood donations as part of the comprehensive

blood safety initiatives through the implementation of effective donor recruitment strategies and activities would maximize the blood donation(17).

Currently the ERCS-NBBS has the following organizational structures dealing with specific activities (17): Blood donor service management, laboratory division, quality control division, data analysis unit and administrative and finance unit.

- Blood Donor Service Management Division is responsible for administering the donation and recruitment of blood to distributing blood and blood components to different hospitals.
- Laboratory Division has a mandate of undertaking the screening of blood from transfusion transmittable infectious diseases.
- Quality Control Division Unit as the name dictates has responsibilities of quality assurance of the equipments and kits that are used for the screening purpose and other related functions.
- Data Analyzing Unit keeps track of every donors information and statistically report the .collected blood to the dedicated bodies such as FMOH,WHO,CDC and the directorate office
- Administration and Finance Unit is concerned with the financial matters to undergo the recurrent activities that demands financial intervention.

Blood transfusions worldwide currently face appealing challenges. Transfusion transmissible infections, such as HIV, Hepatits B virus (HBV), Hepatitis C virus (HCV), Syphilis, and Malaria have provoked a greatly heightened emphasis on safety with inescapable implications for the complexity and cost of providing a transfusion service. One of the biggest challenges to blood safety particularly in Sub-Saharan Africa is accessing safe and adequate quantities of blood and blood products(29).Communities in Africa face several enduring challenges: chronic blood shortages, high prevalence of transfusion transmissible infections (TTI's), lack of national blood transfusion services, recruitment and retention of voluntary nonremunerated donors, family replacement and commercial blood donation. Addressing these challenges would be a central

priority for most blood transfusion services, particularly in Sub-Saharan African countries, to ensure the uninterrupted supply of safe blood and blood products.

According to Tagny (29) serious blood shortages also contribute to an increased risk of HIV and Hepatitis because an inadequate stock of blood forces a reliance on unsafe family or paid donors and increased pressure to issue blood without testing. In 2004, about 1.2 million units of blood were collected from family or paid donors who are considered at high risk for transmitting HIV, Hepatitis B or Hepatitis C. Only 12 sub-Saharan countries[5] have achieved 100 per cent voluntary unpaid blood donation, which is the cornerstone of a safe blood supply.

In principle, safe and sufficient supply of blood and blood products should be secured for all patients requiring transfusion(30). Countries may also formulate a national blood policy and plan, as part of the national health policy, to define how safe blood and blood products will be made available and accessible to address the transfusion needs of its population, including how blood transfusion services will be organized and managed.

Developing countries face considerable obstacles to ensuring a safe blood supply and safe blood transfusions. Because developing countries tend to have inadequate available blood supplies, they depend on family blood donors(31). A family replacement donor is one who gives blood when it is required by a member of the donor's family or community. One disadvantage of this method of blood donation is that patients or their relatives are under intense strain when the patient is admitted to hospital. Being expected to provide replacement donors puts additional responsibility and stress on relatives, and there is undue pressure on family members to give blood, even when they know that donating blood may affect their own health or that they may be potentially at risk of transmitting TTI's(32, 33).

A country's transfusion needs cannot easily be met by relying solely on family replacement donations. The World Health Assembly recommended that reliance on replacement donations should be phased out due to their association with an increased risk of TTI's(34).Meeting the transfusion needs of recipients is challenging because donated blood may not necessarily be

replaced in type or quantity(34). This leaves relatives who cannot find suitable donors with no other option than to seek commercialy remunerated, high-risk blood donors. Blood donated by certain relatives, particularly spouses of women of child-bearing age, can put their wives/partners potentially at risk of producing antibodies to clinically significant antigens that the husband and the developing fetus may have but which the wife lacks. There are increasing concerns about the sustainability of centralized voluntary donor systems and their compatibility with the suboptimal level of healthcare facilities existing in many Sub-Saharan African countries, yet burdening patients' families with the responsibility of finding replacement blood donors will exacerbate poverty and reduce the safety of the blood supply.

The provision of safe and efficacious blood and blood components for transfusion or manufacturing use involves a number of processes, from the selection of blood donors and the collection, processing and testing of blood donations to the testing of patient samples, the issue of compatible blood and its administration to the patient. There is a risk of error in each process in this "transfusion chain" and a failure at any of these stages can have serious implications for the recipients of blood and blood products. Thus, while blood transfusion can be life-saving, there are associated risks, particularly the transmission of blood borne infections(35).

Screening for transfusion-transmissible infections (TTI's) to exclude blood donations at risk of transmitting infection from donors to recipients is a critical part of the process of ensuring that transfusion is as safe as possible. Effective screening for evidence of the presence of the most common and dangerous TTI's can reduce the risk of transmission to very low levels(36). Blood transfusion services should therefore establish efficient systems to ensure that all donated blood is correctly screened for specific TTI's and that only non-reactive blood and blood components are released for clinical and manufacturing use.

A study undertaken to determine the seroprevalence of HIV infection among 1500 blood donors living in the Niger Delta area of Nigeria showed a prevalence of 1.0%. The highest prevalence occurred among commercially remunerated donors(37). Similarly a study of 33,682 and 1259 blood donors screened in 2 tertiary hospitals in Nigeria indicated overall HIV seroprevalence of

7.66% and 0.71%, respectively(37). Two studies to investigate the risk of transfusion-transmissible HIV infection among Malian blood donors indicated prevalences of 2.6%36 and 4.5%, respectively(37). Undetectable HIV infections in blood banks pose a serious threat to public health. In Kenya, blood donations from high school students are preferred over adult samples due to the lower HIV infection prevalence within this population. However a study carried out using Stimmunology, an in vitro lymphocyte stimulation technique,has revealed a significant number of early, preseroconversion HIV carriers both among adult and teenage Kenyan populations(38).

HBV is also the most common cause of serious liver infection in the world. Worldwide, it is estimated that more than 2 billion people have been infected by HBV and 350 million people have chronic infection(39). HBV is highly contagious and transmission of HBV occurs through percutaneous or permucosal routes, infective blood or body fluids introduced at birth, sexual contact,or contaminated needles. Transfusion transmitted HBV infection is increasingly becoming a major mode of transmission of HBV in high-prevalence areas in Sub-Saharan Africa. Rate of positive donations per blood unit collected among Malian blood donors was 13.9%. Hepatitis B surface antigen (HBsAg) -positive donations were significantly higher among donations from replacement donors than those from volunteer donors(39). A cross-sectional study undertaken to determine the seroprevalence of HBsAg among 1410 apparently healthy prospective blood donors in Nigeria observed an overall seroprevalence of 18.6%(38).

A study was conducted by Baye Gelaw(19) to determine the prevalence of HBV, HCV and malaria parasites among healthy adult blood donors in Gondar, Bahirdar, Dessie and Mekele blood banks. Result of the study indicates the overall prevalence of HBV, HCV and malaria parasites were 6.2%, 1.7% and 1% respectively. Magnitude of the prevalence under this study might warrant the introduction of screening of all blood donors for Hepatitis viral markers (HBV and HCV) and should be instituted in parts of the country.

The adoption of screening strategies appropriate to the needs, infrastructure and resources of each country can contribute significantly to improvements in blood safety. In countries where

effective blood screening programmes have been implemented, the risk of transmission of TTI's has been reduced dramatically over the last 20 years(40, 41).

Nevertheless, a significant proportion of donated blood remains unsafe as it is either not screened for all the major TTI's or is not screened within a quality system. Data on blood safety indicators provided in 2007 by Ministry of Health to the WHO Global Database on Blood Safety (GDBS) indicate that, of the 155 countries that reported Performing 100% screening for HIV, only 71 screen in a quality-assured manner(42).Concerted efforts are still required by a substantial number of countries to achieve 100% screening of donated blood for TTI's within quality systems.

2.1 Overview of Data Mining

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the routine (such as super market transaction data, credit card usage records, telephone call details and government statistics) to the more exotic such as image astronomical bodies, molecular databases and medical records. These days interest has grown in the possibility of extracting information from the databases that might be of valuable to the owner of the database. The discipline concerned with the task has become known as Data Mining(43).

It has been estimated that the amount of information in the world doubles every 20 months(44). The size and number of databases probably increases even faster; that is, many scientific, government and corporate information systems are being plagued by the gigantic production of data that are generated and stored routinely, which grow into large databases amounting to giga bytes (and even tera bytes) of data(44).The author further argued that given certain data analysis goal, it has been a common practice to either design a database application on on-line data or use a statistical (or analytical) package on off-line data along with a domain expert to interpret the results. Even if one does not count the problems related with the use of standard statistical packages (such as its limited power for knowledge discovery, the need for trained statisticians and domain experts to apply statistical methods and to refine/interpret results, etc.),one is

required to state the goal and gather relevant data to arrive at that goal. Consequently, there is still strong possibility that some significant and meaningful patterns in the database, waiting to be discovered, are missed(44).

Recent advances in communication technologies, on the one hand, and computer hardware and database technologies, on the other, have made it all the more easy for organizations to collect, store and manipulate massive amounts of data. As stated by Deogan (44) ,these large databases contain potential gold mine of valuable information, but it is beyond human ability to analyze substantial amounts of data and extract meaningful patterns. As the volume of data increases, the proportion of information in which people could understand decreases substantially. The author further stated that given certain data analysis goal, it has been a usual practice to either design a database application on on-line data or use a statistical (or an analytical) package on off-line data along with a domain expert to interpret the result. This traditional method of turning data into knowledge in most application areas, such as marketing, finance, retail, insurance, science, etc., relies on manual analysis and interpretation. Moreover, it demands one or more analysts who become thoroughly familiar with the data and serving as an interface between the data, the users and products. This form of manual probing of a dataset is slow, expensive and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains(45).

Rea (13) wrote that in the last two decades it has been observed a dramatic increase in the amount of information or data being accumulated in electronic format. Having concentrated on the growth of data, it springs a question what to do next with this valuable resource? Indeed, the data contains and reflects activities and facts about the organization. But the data's hidden value, the potential to predict business trends and customer behavior, has largely gone untapped(13).Thus, to provide useful information and knowledge about a business by going beyond the data explicitly stored, the data stored in databases or paper files should be analyzed and interpreted into knowledge. However, statistical theory and practice, which for many years has been the traditional method to study and analyze data, fail when it comes to analyzing large amounts of data(13).

As Han and Kamber (45)stated the major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research(45).The authors also said that data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of functionalities such as data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining).

It is to narrow this gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as Data Mining has emerged in recent years.

Data mining is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, heuristics, data acquisition, data visualization, optimization, information retrieval, high end computing, and others (14, 45).

According to Fayyad and Piatetsky (5), data mining is a process of non-trivial extraction of implicit, previously unknown, potentially useful and actionable information (such as knowledge rules, constraints, regularities) from large amounts of data in databases. This information enables to make critical business decision.

As Berry and Linoff (46) stated, data mining usually makes sense when there is huge amount of data. On account of this reason most of the algorithms developed for data mining purpose requires large volume of data so as to build and train models that are intended to be used for different tasks of data mining such as classification, clustering, and association rule discovery. The rationale behind the need for bulky data is simple and straightforward, small training data

results in unreliable generalizations based on chance patterns. As a result, most data mining tools and algorithms demand large amount of training data (data used for building a model) in order to generate unbiased models(46).

2.2 Data Mining and Knowledge Discovery in Databases (KDD)

Historically, the notion of finding useful patterns from data had been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing. As Piatetsky (48) notes, the term data mining has mostly been used by statisticians, data analysts, and the management information system (MIS) communities. It has also gained popularity in the database field.

KDD is; the nontrivial process of identifying valid, novel, implicit, potentially useful, and ultimately understandable patterns in data(48). Many people treat data mining as a synonym for the phrase Knowledge Discovery in Databases or KDD. On the other way others view data mining as simply an essential step in the process of knowledge discovery in databases. Han and Kamber (45) agree to the second view that KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. More specifically, according to Brachman and An (2000) as quoted in(49), although at the core of the knowledge discovery process, the data mining step usually takes only a small part (estimated at 15% to 25 %) of the overall effort..The data mining component of the KDD process is the application of specific algorithms for extracting patterns from data and heavily relies on known techniques from machine learning, pattern recognition, and statistics.

According to Piatetsky (48),the phrase Knowledge Discovery in Databases(KDD) was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the Artificial Intelligence (AI) and machine learning fields.

Although these fields provide some of the data mining methods, KDD focuses on the overall process of knowledge discovery from data. These focuses of KDD include how the data are

stored and accessed; how algorithms can be scaled to massive data sets and still run efficiently; how results can be interpreted and visualized; and how the overall man-machine interaction can usefully be modeled and supported(5).

By grounds of the popularity of the term 'data-mining' than the term 'Knowledge Discovery in Databases', Han and Kamber(45) inclined to adapt the broader view of data mining functionality, and defined data mining as the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repository. Thus as to the usage of these two phrases, 'data mining' and 'knowledge discovery in databases', the broader view as it has been adapted and defined by Han and Kamber(50) is adapted in this research. The reasons behind this adaption are, being consistent with major data mining studies, use the corresponding experiences, and avoid any confusion between the two phrases, 'data mining' and 'knowledge discovery in databases'.

2.3 Data Mining and Statistical Tools

Analyzing data which is either in a manual or electronic form was a common undertaking. Statisticians have been collecting and analyzing data for ages(51).The computer itself has been used as a tool for data analysis at least some thirty years ago. However, the new era, era of information age has brought some considerable changes to the area of data analysis. As (51)argues many business processes have been computerized which resulted to a significant increase in the amount of available data.

Rea (13) also argues that to detect unusual patterns and explain patterns using statistical models such as linear models, analysts have used statistical analysis systems such as SAS and SPSS.

According to (51) 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. More sophisticated uses of hypothesis testing

include one-way and two-way analysis of variance (ANOVA), where one or two independent factors are tested for affecting another variable, called the “response”. As (51) puts it the hypothesis testing can be a practical tool for supporting a decision-making process, but not for improving our knowledge about the world.

The regression methods (simple linear, multiple linear and non-linear models) represent the more discovery-oriented approach of the classical statistics, because they enable to find the unknown coefficients of mathematical equations relating a dependent variable to its predictors. Regression methods are very efficient in computation but they are limited to use with continuous (numeric) attributes only(51).Moreover, regression methods assume a pre-determined form of functional dependency (e.g. linear) and provide no indication on existence of their functional dependencies in data.

Unfortunately, statistical theory and practice, which for many years has been the traditional method to study and analyze data, fail when it comes to analyzing large amounts of data(52) .A slightly better situation is met with the OLAP (Online Analytical Processing) tools, which can be termed visualization driven since they assist the users in the process of pattern discovery by displaying multi-dimensional data graphically.

Online Analytical Processing (OLAP) is the application of traditional query-and-reporting programs to describe and extract what is in the database. The user forms hypothesis about the data relationships and employees OLAP to verify the hypotheses with queries of the database(53).

Because of 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 to mitigate the drawbacks witnessed with traditional statistical methods(49).

Levin and Zahavi (14) have also stated that until recently, the ability to analyze and understand volume of data lagged far behind the capability, to gather, store and manipulate the data but not any more after the advent of data mining technology. However, it does not mean that data mining has replaced other statistical methods such as OLAP, Regression, etc. Rea (13) 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.

Graetinger (51) also stated that data mining does not replace but rather complements and interlocks with other decision support system capabilities such as query and reporting, on-line analytical processing (OLAP), data visualization and traditional statistical analysis.

2.4 Data Mining Models

Before one attempts to extract useful knowledge from data, it is important to understand the overall approach to be followed. Simply knowing many algorithms used for data analysis is not sufficient for a successful data mining (DM) study. Having a clear description of the process models to be used can gear to the different steps to be followed which helps find new knowledge. Usually the process defines a sequence of steps (with eventual feedback loops) that should be followed to discover knowledge (e.g., patterns) in data(15).

As Tesfaye (54) stated there is a confusion with people in that data mining seems a mere application of software's but it is more than this. In fact it is a process that involves a finite series of steps to process the data prior to mining and post processing steps to evaluate and interpret the modeling results.

The knowledge discovery process consists of a set of processing steps to be followed by practitioners when executing a knowledge discovery study. The model describes procedures that are performed in each of its steps(55).

Since the 1990s, several different KDPs have been developed. The initial efforts were led by academic research but were quickly followed by industry. The first basic structure of the model was proposed by Fayyad (1996) and later improved/modified by others such as Cabana, Anand and Buchner. The process consists of multiple steps that are executed in a sequence. Each subsequent step is initiated upon successful completion of the previous step, and requires the result generated by the previous step as its input. Another common feature of the proposed models is the range of activities covered, which stretches from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results. All the proposed models also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process(56).

Although the models generally give emphasis to independence from specific applications and tools, they can be broadly categorized into those that give attention to the industrial issues and those that do not. However, the academic models, which often are not concerned with industrial issues, can be rather made applicable quite easily in the industrial setting and vice versa(56).
The efforts to establish a KDP model were initiated in academia. In the mid-1990s, when the DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The two process models developed in 1996 and 1998 are the nine-step model by Fayyad et al (1996). and the eight-step model by Anand and Buchner(22).

The Fayyad et al.(1996) KDP model which is developed with the intent to apply in the academic settings consists of nine steps. Cabena (55) argues that, Nevertheless, a number of loops between any two steps are usually executed, but they give no specific details. The model provides a detailed technical description with respect to data analysis but lacks a description of business aspects. However this model has become an important milestone for later models.

Industrial models quickly followed academic models. Several different approaches were undertaken, ranging from models proposed by individuals with extensive industrial experience to models proposed by large industrial consortiums. Two representative industrial models are the

five-step model by Cabena(55), with support from IBM and the industrial six-step CRISP-DM model, developed by a large consortium of European companies which the later become the leading industrial model(55).

The CRISP-DM (CRoss-Industry Standard Process for Data Mining) was first established in the late 1990s by four companies: Integral Solutions Ltd. (a provider of commercial data mining solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company). The development of this process model enjoys strong industrial support(55).

The CRISP-DM has six steps with frequent feedback loops between the subsequent steps. Unlike that of the academic models which are focused with academic settings this model is developed to solve business issues that needs deployment results of the knowledge discovery process.
The development of academic and industrial models has led to the development of hybrid models; models that integrate features of both. One such model is a six-step KDP model developed by Cios(22). It was developed based on the CRISP-DM model by adopting it to academic research.

The KDP model provides more general, research-oriented description of the steps, by introducing a data mining step instead of the modeling step used in the CRISP-DM. The knowledge discovery process model is iterative, and involves numerous steps with many decisions made by the user. This iterative process has been summarized by many researchers and the area professionals. Most of them agree that knowledge discovery process starts with a clear definition of the business problem or, equivalently, understanding of the application domain(22, 23, 57).

The knowledge discovery process model for data mining generates an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks and relationships between these tasks. At this description level, it is not possible to identify all relationships because there is a feedback communication between each phase of the life cycle.

Essentially, relationships could exist between any data mining tasks depending on the goals, the background and interest of the user and most importantly on the data(57).The life cycle of a data mining project both in the CRISP-DM and KDD model consists of six phases. As there is always a dynamic communication, the sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase, for which phase or which particular task of a phase, has to be performed next(57). The detailed explanation of the phases followed in each model is presented below.

Figure 2.1 shows descriptions of the six steps of the CRISP-DM process model(58).

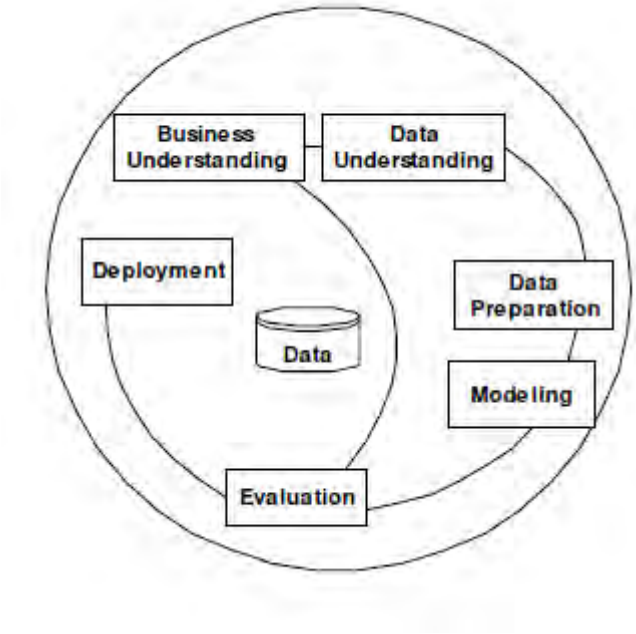

Figure 2.1: The CRISP-DM process model

Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

Evaluation

At this stage of the project, one has to build a model (or models) that appear(s) to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.

2.5 Data Mining Techniques

According to Berry and Linoff (46),having an in depth knowledge and understanding of different data mining techniques is indispensable for the following reasons:

- In order to make use of and take the advantage of a specific technique, it is important to know the details of each technique.
- To find out the best applicable technique for the problem at hand.
- To know the advantages and disadvantages of a technique.

It is evident that no one technique is applicably suited to all data mining problems. Determining the best technique that fits to the specific data mining problem and familiarizing with the available techniques is extremely essential. According to(56), the most commonly used data mining techniques are: Decision tree, neural networks, genetic algorithms, nearest neighbor method and rule induction.

According to Levin and zahavi (14) data mining techniques can be categorized into two major application groups: Predictive modeling and descriptive modeling. In each of these applications, data mining differs in the approach taken to solve problems. Each application is usually geared in solving a particular type of problem. That is, a specific algorithm is favored over others depending what the problem posed by the data miner.

According to Levin (1999), Han and Kamber (45) in predictive modeling tasks, one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models such as classification, regression and other AI-based models. Predictive models are

built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results.

On the other hand, descriptive models belong to the realm of unsupervised learning; it is called unsupervised learning since there are no already known results to guide the algorithm. Such models interrogate the database to identify patterns and relationships in the data. Clustering ,segmentation and visualization methods, among others, belong to this family of descriptive models(14).As Han and Kamber (45) stated, in this unsupervised learning users may sometimes have no idea which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus, it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations of applications.

2.5.1 Predictive modeling

Prediction is arguably the strongest goal of data mining. The aim is to build a model that can permit the value of one variable to be predicted from the known values of other variables. Classification and regression are two good examples of data analysis that can be used to extract models describing important data classes or to predict future data trends(45).

Classification problems, as stated by Deogan (59) , aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. However, the quality of the discovered knowledge is heavily dependent on the algorithms used to analyze the data.

Classification methods create classes by examining already classified cases and inductively finding the pattern (or rule) typical to each class. Data mining uses machine – learning methods such as decision trees, neural networks and bayes to classify objects based on a dependent variable.

According to Han and Kamber (45) data classification is a two-step process. In the first step, a model is constructed by analyzing database tuples described by the attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. In the second step the model is used for classification. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. Then the accuracy of a model on a given test set is evaluated.

As Rea (13) stated, once classes are defined the system should infer rules that govern the classification and therefore should be able to find the description of each class. The writer further argued that the descriptions should only refer to the predicting attributes of the training set so that the positive examples should satisfy the description and none of the negative. A rule is said to be correct if its description covers all the positive examples and none of the negative examples of a class. Basically, in classification tasks, the system, given a case or tuple with certain known attribute values should be able to predict to which class this case belongs to.

Although the choice of techniques suitable for classification tasks seems to be strongly dependent on the application, the data mining techniques that are frequently employed for classification tasks are the Decision trees, Bayes and Neural networks(60).Since this study is a classification task that makes use of different classification techniques, the decision trees and the neural networks are discussed below.

2.5.1.2 Decision tree

One of the most commonly used data mining techniques for classification tasks are decision trees. Decision trees are simple knowledge representation and they classify examples to a finite number of classes. In decision tree induction, the nodes of the tree are labeled with attribute

names, the edges of the tree are labeled with possible values for the attributes and the leaves of the tree generate decision tree from a given set of attribute-value tuples.

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

A decision tree is constructed by repeatedly causing a tree construction algorithm in each generated node of the tree(61). The classification is performed separately for each leaf, which represents a conjunction of attribute valued in a rule(51).
Structurally, decision trees consist of two types of nodes; non-terminal (intermediate) and terminal (leaf). The former correspond to questions asked about the characteristic features of the diagnosed case. Terminal nodes, on the other hand generate a decision(62).
A typical decision tree is shown in Figure 2.2 below.

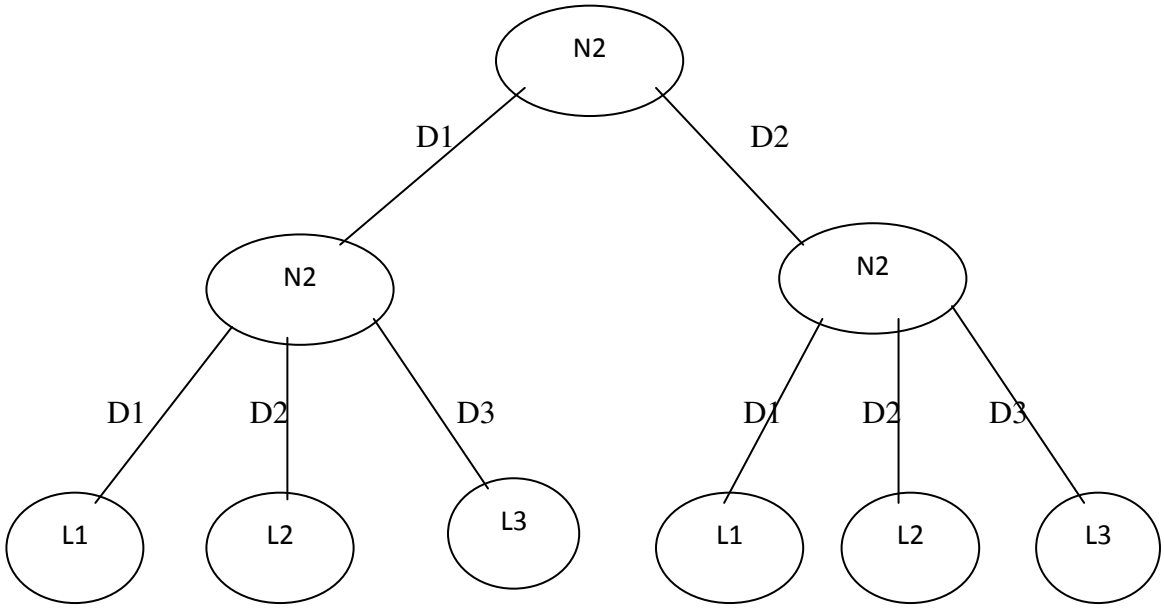

Figure 2.2.: A Decision Tree with Decision (Ni) and Leaf (Li) nodes, and decisions (Di)(9)

The very advantage of decision trees is that they can handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations and the explosion of predictor variables inherent in neural nets(23).

The commonly agreed drawbacks for which the decision tress are criticized is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In addition, all splits are made sequentially, so each split is dependent on its predecessor(23).

2.5.1.2 Neural networks

The leading models in the AI based knowledge discovery are Neural Networks (NN) models. NN is a biologically inspired model which tries to mimic the performance of the network or neurons, or nerve cells, in the human brain. Expressed mathematically, a NN model is made up of a collection of processing units (neurons, nodes), connected by means of branches, each characterized by a weight representing the strength of the connection between the neurons. A typical NN contains several input nodes connected to one or more output nodes, through an intermediate set of hidden nodes(63).

The structure of neural network is very similar to the structure of the neurons in the human brain. All of the processing of a neural network is carried out by this set of neurons or units. Each neuron is a separate communication device, doing its own relatively simple job. A unit's function is simply to receive input from other units and, as a function of the inputs it receives, to compute an output value, which it sends to other units. The system is inherently parallel in that many units can carry out their computations at the same time(63).

According to Frohlich (63), In neural networks, neurons are grouped in layers, often classified as input, hidden, and output layers.Inputs layer is a processing element that receives the input to the neural network and hidden layers are processing elements between a neural network's input layer

and its output layer.On the other hand, output layer is the processing element that produces neural network's output.

There could be a number of input, hidden and output neurons in each corresponding layer. For example, in the following neural network (figure 2.3), there are three inputs, three hidden and three output neurons. In fact, the network consists of one input, hidden and output layer.

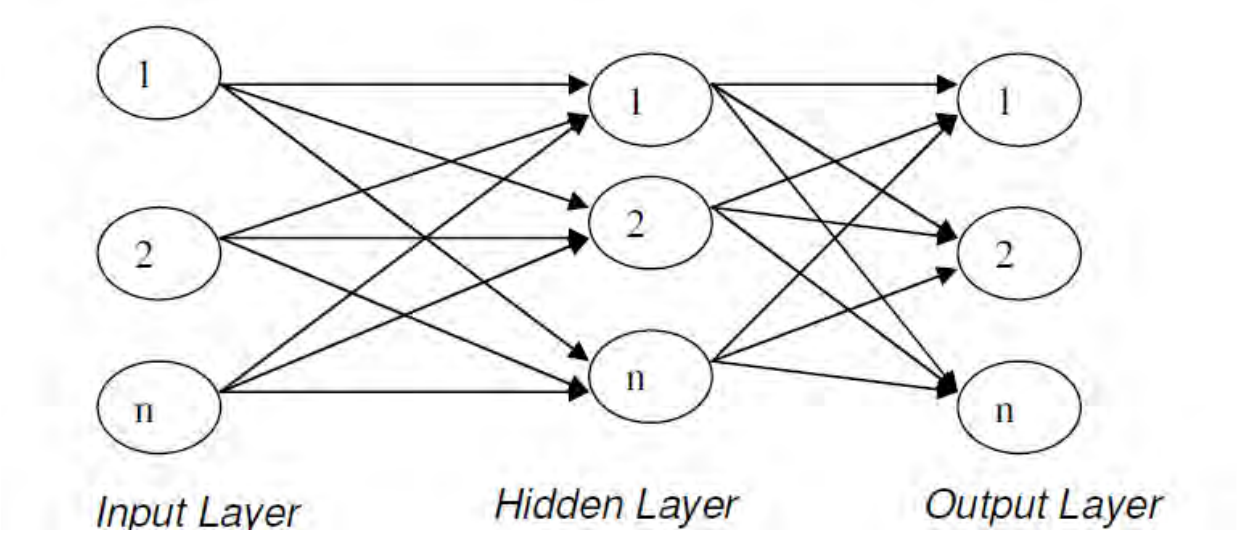

Figure 2.3: A simple neural network

NN have become of particular interest in data mining because they offer a means for efficiently modeling large and complex problems in which there are hundreds of independent variables that have many interactions (63).

2.5.2 Descriptive modeling

Description involves using some variables or fields in the database and focuses on finding human-interpretable patterns describing the data(5).One example of such models is clustering (or segmentation) algorithm.

2.5.2.1 Clustering models

Han and Kamber (45) stated the process of grouping a set of physical or abstract objects into class of similar objects is called clustering. Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. It is mapping a data item into one of several clusters which are not pre-specified but are determined from the data. Clusters are formed by finding natural groupings of data items based on similarity matrices, proximity considerations and probability measures(45).

Two Crows Corporation (23) mentioned that the goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. The categories (clusters) can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories.

According to Han and Kamber (45), each cluster that is formed can be viewed as a class of objects from which rules can be derived.

As Rea (13) points out there are a number of approaches for mining clusters. They are in general categorized in to partitioning and hierarchical clustering.

2.5.2.1.1 Partitioning clustering algorithm

The partitioning clustering algorithms group data into un-nested and non-overlapping groups that usually optimize a clustering(64).Levin and Zahavi (14) argues that perhaps the most common of all automatic partitioning clustering is the K-means algorithm, which assigns observations to one of K classes to minimize the within-cluster-sums-of-squares. Also worth mentioning are the Judgmental-based or manual segmentation methods which are still very popular in direct marketing applications to carve up a customers list into homogeneous segments.

2.5.2.1.2 Hierarchical clustering algorithm

Hierarchical clustering algorithm groups data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering(64). In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach.

2.5.2.2 Association rule discovery

According to (65) discovery is the process of looking in a database to find hidden patterns without a predetermined idea or hypothesis about what the patterns may be. In other words, the program takes the initiative in finding what the interesting patterns are, without the user thinking of the relevant questions first. In large databases, there are so many patterns that the user can never practically think of the right questions to ask. The key issue here is the richness of the patterns that can be expressed and discovered and the quality of the information delivered. This in turn determines the power and usefulness of the discovery technique.

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. The rules are given in the form: if item A is part of an event, then X% of the time item B is also part of the event. The rules are written as A→B,where A is called the antecedent or left-hand side (LHS), and B is called the consequent or righthand side (RHS). More formally, association rules are of the form A→B, that is ,

(A_1, - - - , $A_m \rightarrow B_1$, - - - - - B_n) Where A_i (for i =1, - - - m} and B_j (for j = 1, - - - - n } are ttribute-value pairs.

The associations rule A→B is interpreted as database tuples that satisfy the condition in A are also likely to satisfy the condition in B.

Support and confidence are the probability measures, introduced to assess associations in the database. The support (or prevalence) of a rule is the proportion of observations that contain the item or item set of the rule. It is also known as the coverage of the rule.

As stated by Witten and Frank (66), an item is an attribute value pair. The confidence is the conditional probability of B given A, P (B/A). A rule is "interesting" if the conditional probability P (B/A) is significantly different than P (B). Confidence of the rule measures the rule's accuracy.

Association algorithms find these rules by doing the equivalent of sorting the data while counting occurrences so that they can calculate confidence and support. The efficiency with which they can do this is one of the differentiators among algorithms. One should be able to evaluate rules using different techniques especially because of the combinatorial explosion that results in enormous number of rules(66).

As written by Han and Kamber (45) ,association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Rules that satisfy both a minimum support threshold and a minimum confidence threshold are strong rules.

A good example of the use of associations is the analysis of the claim forms submitted by patients to a medical insurance company. By defining the set of items to be the collection of all medical procedures that can be performed on a patient and the records to correspond to each claim form, the application can find, using the association function, relationships among medical procedures that are often performed together(67).

In specific situations the association discovery components can help one to:

- Manage existing customers. Determine response propensities by segmenting customers on purchase patterns and attributes.
- Use knowledge of customer segment attributes to recommend items or actions that might appeal to each segment.

- Acquire new customers. Analyze purchase pattern and attribute data from an outside source to develop customer segmentation models. Then "acquire" new customers whose characteristics resemble those of your best customers by offering them targeted products and services.
- Detect patterns of potentially harmful behavior.Detect patterns of events or behavior that can help identify the potential of bioterrorist attacks and infrastructure intrusions.
- Spot fraud, waste and abuse. Detect patterns of fraudulent and abusive behavior so you can take steps to prevent future occurrences.
- Improve Web site navigation. Make it easier for people to make Web-based purchases by enhancing site navigation and how items are presented.
- Medical diagnosis/research. Identify telltale symptoms to aid in effective diagnosis.

2.5.2.3 Sequence discovery

Sequence discoveries are association rules with time dimensions. A sequential pattern is an association between sets of items, in which some temporal properties between items in each set and between sets are satisfied. In particular, items in a set have the same temporal reference(14).

As Trybula (53) described, sequential patterns are identified in a technique for predicting future activities based on observing trends over a period of time. It is based on the fact that previous activities have the potential for indicating future activities.

Two Crows Corporation (23) points out that association or sequence rules are not actually rules, but rather descriptions of relationships in a particular database. There is no formal testing of models on other data to increase the predictive power of these rules rather there is an implicit assumption that the past behavior will continue in the future.

2.6 Related Works

These days numerous health care organizations and research institutions worldwide are investigating the application of data mining techniques in order to improve the effectiveness and efficiency of health related services.

2.6.1 Clinical data mining

Current trends in medical decision making show awareness of the need to introduce formal reasoning, as well as intelligent data analysis techniques in the extraction of knowledge, regularities, trends and representative cases from patient data stored in medical records (68). According to Larvac (6), machine-learning methods have been applied to a variety of medical domains in order to improve medical decision making.

Prather et al.(3) has undergone a data mining project at Duck University Medical Center using an extensive clinical database of obstetrical patients to identify factors that contribute to prenatal outcomes. The objective of this knowledge discovery effort was to identify factors that can improve the quality and cost effectiveness of prenatal care. The production system database identified for mining was the computer-based patient record system known as "The Medical Record", or TMR. TMR is a comprehensive longitudinal clinical patient record system (CPRS) developed at Duke University over the last 25 years(69).For their work, the specific database selected for the data mining project was the prenatal database used by the Department of Obstetrics and Gynecology at Duke University Medical Center. For the purpose of the initial study, the researchers created a sample two-year dataset (1993-1994) from the data warehouse. Exploratory factor analysis was selected for data mining. The statistical software used to conduct the factor analysis was SPSS for windows version 5.0(69).

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

2.6.2 Knowledge discovery from mortality databases

Last (51) have applied the process of knowledge discovery to a dataset of 33,134 mortality records extracted from the Israeli Ministry of Health mortality database. The data set analyzed by these researchers includes the records of all Israeli citizens who passed away in the year 1993.The purpose of this study was to identify the leading causes of death and the association of various factors with certain diseases by applying data mining techniques. The death cause (medical diagnosis) of each person was defined by the International 6-digit code (ICD-9-CM).To carry out the project successfully, a significant amount of data preparation was performed. The information-theoretic approach to data mining was the approach that has been used by these researchers. According to the authors, the data mining process has resulted in selection and scoring of the most important input attributes discretization of continuous features, rule extraction, and calculation of data reliability. Finally, the following main results were obtained:

The automated data cleaning procedure has revealed outliners in most attributes of the database. Traditionally, the mortality data has been analyzed with respect to age, gender, and ethnic origin. However, the results achieved by their study suggested that time of year are more important factor for death than place of birth and/or ethnic origin. Rules defining high-risk and low-risk groups (with respect specific causes) have been extracted and scored by the information theoretic network. The researchers argue, these rules can be used for determining priorities in the health care budget. They may be also valuable for insurance companies and other commercial institutions. Most unreliable records in the database contain lowly probable information. The authors further proposed that this information should be checked by medical experts and possibly compared to the manual source. This comparison can lead to correcting the data in the original database.

Lloyd-Williams (70) had also analyzed datasets extracted from the World Health Organization's Health for All (HFA) database using a data mining approach. During the selection process, mortality data based on the following conditions was extracted by the researchers from the HFA database: life expectancy at birth; probability of dying before five years of age; infant mortality;

post-neonatal mortality; standardized death rate (SDR) for circulatory diseases; SDR for malignant neoplasm; SDR for external causes of injury and poisoning; SDR for suicide and self-inflicted injury. Data was extracted for 39 European Countries, and then converted into a format acceptable to the software used for that particular project.

An underlying aim of the study, as noted by the author, was to track changes in the data that may have occurred over the years for the same samples of countries in order to examine whether any patterns identified remained consistent over time. The extracted data was analyzed by custom written Kohonen self-organizing map software in order to identify possible groupings. Standard statistical techniques were used to evaluate the validity of the groupings(70).

Preliminary work of Lloyd-William's study resulted into two groups for clusters of countries in each year being apparent. In addition to the geographical division, the classification also appeared to reflect differences in wealth. Countries in the first of the groups were relatively poor; whereas countries in the second of the groups were relatively wealthy. Lloyd-Williams (1996) reported that the observation that the classification appeared to reflect two different GNP groups suggested that GNP could be inter-related with the health indicators. In order to further explore this possibility, the author calculated coefficient of correlation between GNP and all the seven HFA indicators he has used in the initial analysis. Results obtained indicated that GNP is strongly and positively correlated with life expectancy, and strongly but negatively correlated with the SDR for diseases of the circulatory system.

2.6.3 Mining pediatric primary case database

Dons and Wallace (20) have also applied data mining techniques to mine association rules from a pediatric primary care decision support system. According to the authors, the purpose of their study was to apply an unsupervised data mining algorithm to a database containing data collected at the point of care for clinical decision support. They took the data set from the Child Health

Improvement Program (CHIP), a preventive services tracking and reminder system in use at the University of North Carolina. The workers used the unsupervised data mining (pattern discovery) algorithm to extract 2nd and 3rd order association rules from the data. As a result of the data mining process, the algorithm, which the authors have used, discovered 16 2nd order associations and 103 3rd order associations. The authors have also identified that the 3rd order associations contained no new information. However, the 2nd order associations demonstrated a covariance among a range of health-risk behaviors. Additionally, the algorithm discovered that both tobacco smoke exposure and chronic cardiopulmonary disease are associated with failure on developmental screens. Summarizing their results, (20) stated that the discovery of a direct association between cardiopulmonary disease (e.g., asthma) and developmental delay among otherwise healthy children was a novel discovery. However, the literature shows a high covariance among a range of health risks that may explain the coexistence of these problems in impoverished families.

2.6.4 Mining HIV/AIDS database

Abraham (71) conducted a research on Application of Data Mining Technology to identify determinant risk factors of HIV infection and to find their association rules: the case of Center for Disease Control and Prevention(CDC). Weka software was used to extract the hidden patterns among the variables under the study. The researcher argued one of the important findings noticed under the study was a new insight about risk feeling of the clients and HIV test result. According to the researcher previously it was known that the clients whose reason for test is plan for future are associated with HIV-negative class. This truth has also been verified with experiment too. However, the experiment disclosed that people whose reason for test is having risk, suspect or symptoms is also associated to HIV-negative result with promising evidence. The researcher further noticed this was previously hidden information that domain experts were impressed to hear about. Accordingly a client who has risky perception of oneself has a better chance to be uninfected.

2.6.5 Application of data mining in blood dataset

Research conducted by Santhanam and Shyam (10) on application of CART algorithm in Blood Donors Classification argued that the availability of blood in blood banks is a critical and important aspect in a healthcare system. Blood banks are typically based on a healthy person voluntarily donating blood and is used for transfusions or made into medications. The ability to identify regular blood donors enables blood banks and voluntary organizations to plan systematically for organizing blood donation camps in an effective manner. The researchers identified the blood donation behavior using the classification algorithms of data mining. The analysis had been carried out using a standard blood transfusion dataset and using the CART decision tree algorithm. The CART derived model along with the extended definition for identifying regular voluntary donors provided a good classification accuracy based model (10).

A research was conducted by Baye Gelaw and Yohans Mengistu (19) to determine the prevalence of HBV,HCV and malaria parasites among blood donors in Amhara and Tigray Regional state. The researchers collected blood samples using cross sectional survey from blood donors in northern part of Ethiopia. The socio demographic characteristics of blood donors were assessed using structural questionnaire. The collected blood samples were screened for HBV, HCV and malaria parasites. Their result show that the prevalence of HBV, HCV and Malaria parasites were 6.2%,1.7% and 1 % respectively. a similar research was conducted to determine the seroprevalence of HIV,HBV,HCV and syphilis infections among blood donors at Gondar University teaching hospital north western Ethiopia. A retrospective analysis of consecutive blood donors' records covering the period between January 2003 and December 2007 was conducted. Logistic regression analysis was used to determine risk factors associated with HIV, HBV, HCV and Syphilis infections. The researchers findings shows from the total of 6361 consecutive blood donors, 607 (9.5%) had serological evidence of infection with at least one pathogen and 50 (0.8%) had multiple infections. The overall seroprevalence of HIV, HBV, HCV and Syphilis was 3.8%, 4.7%, 0.7%, and 1.3% respectively. Among those with multiple infections, the most common combinations were HIV - Syphilis 19 (38%) and HIV - HBV 17 (34%). The seropositivity of HIV was significantly increased among female blood donors, first

time donors, housewives, merchants, soldiers, drivers and construction workers. Significantly increased HBV seropositivity was observed among farmers, first time donors and age groups of 26 - 35 and 36 - 45 years. Similarly, the seroprevalence of Syphilis was significantly increased among daily labourers and construction workers. Statistically significant association was observed between syphilis and HIV infections, and HCV and HIV infections. Moreover, significantly declining trends of HIV, HCV and Syphilis seropositivity were observed over the study period.From the above two reviewed literature's one can learn the methods and tools employed to analyze the result are good enough in showing statistical associations and the prevalence of the infectious diseases but the hidden patterns and knowledge's remain untapped.

Apart from the above researches as to the knowledge of the researcher no study was done at the National Blood Bank of Addis Ababa in the same or different techniques and methodologies to apply data mining technology. Hence it is the aim of this research to apply data mining techniques in order to identify donors that are less susceptible to the TTI's diseases and determine the magnitude and significance of the diseases.

CHAPTER THREE

DATA MINING TECHNIQUES

In data mining there are various techniques followed in order to achieve a given data mining goal. Since this is a study which attempts to build a model that predicts safe group blood donors, there is a need of discussing the techniques used in classification and building a data mining model. For this reason the detailed discussion of the techniques is followed in each sub section.

Witten and Frank(66) argues there exist many classification algorithms inside the WEKA system. WEKA (Waikato Environment for Knowledge Analysis) is a group of machine learning algorithms and data processing tools implemented in Java in 1993. It is developed as support for the whole process of experimental data mining such as preparation of input data, statistical evaluation of learning schemes, visualization of input data and the result of learning and used for education, research and applications. Its main features are 49 data preprocessing tools, 76 classification/regression/MLP algorithms, 8 clustering algorithms, 15 attribute/subset evaluators + 10 search algorithms for feature selection, 3 algorithms for finding association rules, 3 graphical user interfaces such as explorer (exploratory data analysis), experimenter (experimental environment) and knowledge flow (new process model interface)(66).

The underlying reason in selecting and discussing the selected techniques goes in addressing the predefined research objectives stated in section one, which is; predicting the seroprevalence of TTI's. Being this the rationale it necessitates justifying models and experiments to be carried out in the course of the knowledge discovery process. For the purpose of building a model different classifier algorithms were employed. Details of each model with subsequent parameters used are presented in section 3.1 and 3.2.

3.1. Classification Model Techniques

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known)(45).
For the classification purpose, Decision tree (J48) to generate rules sets and Bayes (Naives Bayes) to predict class membership probabilities were implemented.

3.1.1. J48 decision tree algorithm

One of the decision tree algorithms used for this study was the J48 decision tree algorithm which is the successor of ID3(Iterative Dichotomiser,) C4.5 .J48 Decision tree is a popular utility that involves decision based classification and adaptive learning over a training set(72).Whitten and Frank(66) further stated J48 algorithm of decision tree technique is one of classification and prediction algorithms which support both numeric and nominal predicators and nominal class attribute values.

The J48 algorithm(73), is the WEKA implementation of the C4.5 top-down decision tree learner proposed by Quinlan in 1993. The algorithm uses the greedy technique and is a variant of ID3, which determines at each step the most predictive attribute, and splits a node based on this attribute. It deals with numeric attributes by determining where thresholds for decision splits should be placed.

An important feature of J48 is a facility of generating outputs both in tree form and rule sets. Graphically it displays the classification process of a given input for given output class labels. Rule sets are generally easier to understand since each rule describes a specific context associated with a class and also shows the hierarchy of the determinant factors or attributes (45).
Decision tree algorithm (44) take inputs, data partition, D, which is a set of training tuples and their associated class labels, attribute_list, the set of candidate attributes and attribute_selection_method, a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes. This criterion consists of splitting attribute and, possibly,

either a split point or splitting subset. The output of the algorithm is a decision tree. Given training tuples the algorithm followed by a decision tree is as followed (45).

1. *Create a node N;*
2. *if tuples in D are all of the same class, c then*
3. *return N as a leaf node labelled with the class C_i*
4. *if attribute_list is empty then*
5. *return N as a leaf node labelled with the majority class in D;// Majority voting*
6. *apply **Attribute_selection_method (D, attribute_list)** to find the "best" splitting_criterion;*
7. *label node N with splitting_criterion*
8. *if splitting-attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees*
9. *attribute_list <-attribute_list-splitting-attribute; // remove splitting-attribute*
10. *for each outcome j of splitting_criterion // partition the tuples and grow subtrees for each partition*
11. *let D_j be the set of data tuples in D satisfying outcome j; // a partition*
12. *if D_j is empty then*
13. *attach a leaf labeled with the majority class in D to nod N;*
14. *else attach the node returned by Generate_decision_tree (Dj, attribute_list) to node N; endfor*
15. *return N;*

Attribute selection method specifies a heuristic procedure for selecting the attribute that best discriminates the given tuples according to class. The process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class).

The process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class). The entropy method of attribute selection is to choose to split on the attribute that gives the greatest reduction in (average) entropy, i.e. the one that maximizes the value of information gain. At any stage of this process, splitting on any attribute has the property that the average entropy of the resulting subsets will be less than (or occasionally equal to) that of the previous training set(74).

This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by:

$$Info\ (D) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Where pi is the probability that an arbitrary tuple in D; belongs to class Ci and is estimated by |Ci, D| / | D |. A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. At this point, the information we have is based only on the proportions of tuples of each class. Info(D) is also known as the entropy of D. Suppose we were to partition the tuples in database D on some attribute A having V distinct values, {a1, a2, . . ., av} as observed from the training data. If A is discrete-valued, these values correspond directly to the V outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, {D1, D2… DV }; where Dj contains those tuples in D that have outcome aj of A. These partitions would correspond to the branches grown from node N. Hypothetically ; we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). The amount of information we would still need (after the partitioning) in order to arrive at an exact classification is measured by:

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} * Info(D_j)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. InfoA (D) is the expected information required to classify a tuple from D based on the

partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$Gain_{(A)} = Info_{(D)} - Info_A(D)$$

Gain (A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain (A)), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum Info A(D)).

Besides to this, the pruned tree in decision tree has a hierarchy in that the most significant variable that is used to discriminate the records is located at the top. It maximizes computational efficiency as well as classification accuracy. The process of pruning (post-pruning) traditionally begins from the bottom of the tree (at the child leaves), and propagates upwards. J48 algorithm recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible.

The overlying principle of pruning is to compare the amount of error that a decision tree would suffer before and after each possible prune, and to then decide accordingly to maximally avoid error. The metric used to describe possible error, denoted error estimate (E), is calculated as

Where 'E' is Err $E= (e+1)/ (N + m)$ ssified examples at the given node, 'N' is examples that reach the given node, and 'm' is all training examples(73).

Applying pruning methods to a tree typically results in reducing the size of the tree to avoid unnecessary complexity (produces fewer, more easily and interpretable results) and to avoid over-fitting of the data set when classifying new data that means improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (73).

In Weka J48 classifier, lowering the confidence factor decreases the amount of post-pruning since the effectiveness is labeled by the confidence factor. Post-pruning in the C4.5 algorithm is the process of evaluating the decision error (estimated percent misclassifications) at each decision junction and propagating this error up the tree (73). At each junction, the algorithm compares the weighted error of each child node versus and Misclassification error (if the child nodes were deleted and the decision nodes were assigned the class label of the majority class).

3.1.2 Naïve Bayes algorithm

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class(75). Naive Bayes is a type of supervised-learning module that contains examples of the input-target mapping the model tries to learn. Such models make predictions about new data based on the examination of previous data.The Naive Bayes algorithm uses the mathematics of Bayes' Theorem to make its predictions(75).

Bayes' Theorem is about conditional probabilities. It states that the probability of a particular predicted event, given the evidence in this instance, is computed from three other numbers: the probability of that prediction in similar situations in general, ignoring the specific evidence (this is called the prior probability); times the probability of seeing the evidence we have here, given that the particular prediction is correct; divided by the sum, for each possible prediction (including the present one), of a similar product for that prediction (that is, the probability of that prediction in general, times the probability of seeing the current evidence given that possible prediction) (45).

A simplifying assumption (the "naive" part) is that the probability of the combined pieces of evidence, given this prediction, is simply the product of the probabilities of the individual pieces of evidence, given this prediction. The assumption is true when the pieces of evidence work independently of one another, without mutual interference. In other cases, the assumption merely approximates the true value. In practice, the approximation usually does not degrade the model's predictive accuracy much, and it makes the difference between a computationally feasible algorithm and an intractable one (75).

The Bayesian Learning Algorithms combine training data with a priori knowledge to get the a posteriori probability of a hypothesis. So it is possible to figure out the most probable hypothesis according to the training data. The basis for all Bayesian Learning Algorithms is the Bayes Rule.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

P(h) = prior probability of hypothesis h

P(D) = prior probability of training data D

P(h|D) = probability of h given D

P(D|h) = probability of D given h One of the algorithms that assume the Bayesian conditional probabilistic in predicting the class member ship and used under this study is the Naive Bayes Algorithm. Up to now it has been surprising predicting which hypothesis is the most probable for a given dataset. But the question what Naive Bayesian classifiers are about is which classification is the most probable for this new instance if we have a look at the training data. For example an instance of Occurrence of disease=(age, sex, occupation, region) could be (less than 24, M, Private Worker, Addis Ababa). A Naive Bayes System could calculate occurrence of disease values for the following three classifications “safe” and “unsafe” according to the

available training data. Then the classifier with the biggest value rules the hypothesis. Where by the conditional independence of the attributes of the instances is required for the use of Naive Bayesian classifiers.

Naive_Bayes_Learn (examples)

for each target value vj

estimate P(vj) for each attribute value ai of each attribute a

estimate P(ai | vj)

The question of how it looks like when brought into a formula is?

Let X be a set of instances xi = (a1,a2,…,an) and V be a set of classifications vj

for each target value vj

estimate P(vj) for each attribute value ai of each attribute a estimate P(ai | vj)

$$v = \max_{vj \in V} P(v_j) \prod_{a_i \in x} P(a_i \mid v_j)$$

Classify_New_Instance (x)

Implementation of Naïve Bayes Classifier in Weka class use estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an UpdateableClassifier (which in typical usage are initialized with zero training instances) -- if one needs the UpdateableClassifier functionality, it is possible to use the NaiveBayesUpdateable classifier. The NaiveBayesUpdateable classifier uses a default precision of 0.1 for numeric attributes when buildClassifier is called with zero training instances.

3.2 Smote

Smote Synthetic Minority Over-sampling Technique (Smote) is an over-sampling method(76). Its main principle is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the over fitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.
On the other hand Kowalczyk (77) , class clusters could not be well defined since some minority class examples might be invading the majority class space. This situation can occur when interpolating minority class examples and can expand the minority class clusters, introducing artificial minority class examples too deeply in the majority class space.

In Smote the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours . Depending upon the amount of over sampling required, neighbours from the k nearest neighbours are randomly chosen(77) .

The detailed algorithm for smote minority class sampling looks as follows:

Algorithm SM OTE (T, N, k)

Input: Number of minority class samples T;AmountofSMOTEN%; Number of nearest neighbors k

Output: (N/100) T synthetic minority class samples*

1. *(* If N i s le ss than 100%, randomize the minority class samples as only a random percen t of them will be SMOTE d. *)*
2. *if N<100*
3. *then Randomize the T minority class samples*
4. *T =(N/100)* T*
5. *N =100*
6. *endif*
7. *N =(int)(N/100)(* The amount of SMOTE i s assumed to be in integral multiples of 100. *)*

8. *k = Number of nearest neighbors*
9. *numattrs= Number of attributes*
10. *Sample[ ][ ]: array for original minority class samples*
11. *newindex: keeps acount of number of synthetic samples generated, initialized to 0*
12. *Synthetic[][]:arrayforsyntheticsamples (* Compute k nearest neighbors for each minority class sample only. *)*
13. *for i← 1 to T*
14. *Compute k nearest neighbors for i, and save the indices in the nnarray*
15. *Populate(N, i, nnarray)*
16. *endfor*

Populate (N, i, nnarray)(Fu nc ti on to g e n e rate t he synthe ti c s ample s . *)*

17. *while N =0*
18. *Choose a random number between 1 and k,call it nn. This step chooses one of the k nearest neighbors of i.*
19. *for attr← 1 to numattrs*
20. *Compute: dif =Sample[nnarray[nn]][attr]- Sample[i][attr]*
21. *Compute: gap= random number between 0 and 1*
22. *Synthetic[newindex][attr]=Sample[i][attr]+gap* dif*
23. *endfor*
24. *newindex++*
25. *N =N- 1*
26. *end while*
27. *return (* End of Populate . *)*

End of Pseudo-Code.

As can be understood in Smote the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours(77). Depending upon the amount of over sampling required, neighbours from the k nearest neighbours are randomly chosen.

3.3 Validation Techniques (Test Options)

For the experimental setup, the original blood datasets were converted to ARFF (Attribute Relation File Format) as this is the suitable input file format for the WEKA system; Subsequently, all the identified algorithms were tested to each blood dataset with the option of using 10-fold cross-validation (the classifier evaluated using the number of folds that are entered in the folds text field).10-fold cross validation used default parameters in the experiment and a standard way of predicting the error rate(73). The k test instances have to be drawn off the training set. The bigger this set is, the more realistic is the estimate of the true error, but the less data is left over for use when training. A static division of the entire set of instances into training and test set may not be representative. Because of this reason the so called cross-validation approach is applied. In cross-validation, the data is partitioned into a fixed number t of disjoint folds which are about the same size. A classifier is built t times; using t - 1 fold for training and one fold for testing.

At the end, every instance has been used once for testing. The test sets are independent, but the training sets which overlap are not independent of each other. The results of the evaluation are averaged over all t runs. The default value for t is ten. Usually, a process called stratification is used in conjunction with cross-validation. Stratification ensures that in each fold the original class distribution is maintained. This improves the stability of the evaluation results.

The default k value (10-fold cross validation) was chosen and used for building the model and test the performance of the model. In this test option the accuracy estimate is the overall number of correct classifications from the k iteration divided by the total number of samples, which is k.

After deciding the values of the parameters the algorithm was run to start building the model. To assure the accuracy of the model, test data was provided to the model to predict for the unknown class value of the blood datasets.

3.5 Evaluation Techniques

Accuracy is the basic measure, which computes, the percent of correctly classified instances in the test set. Accuracy of a test compares how close a new test value is to a value predicted by if…then rules (73). To classify a test example, the rule that matches it best determines the example's class membership. An accuracy test is defined as: Accuracy= (True Positive rate/Total number of test samples)*100%.

When the confusion matrix has only two outcomes (positive and negative) of a test are possible, three evaluation criteria can be used for measuring the effectiveness of the generated rules (73). There are four possibilities, as shown in Table 3.2.

Table 3.1 Probable result of the test set

	Test result Positive	Test result negative
Hypothesis positive	True Positive	False Negative
Hypothesis negative	False Positive	True Negative

Where true positive, indicates the number of correct positive predictions (classifications); true negative is the number of correct negative predictions; false positive is the number of incorrect positive predictions; and false negative is the number of incorrect negative predictions (51). The four accuracy measures are:-

1. Sensitivity = (TP/Hypothesis Positive)*100%= (TP/TP+FN)*100% i.e. the capacity of a test to be positive when the condition is in fact present or how many of the positive test examples are recognized.

2. Specificity= (TN/Hypothesis Negative)*100%= (TN/FP+TN)*100% i.e. the capacity of a test to be negative when the condition is in fact not present or how many of the negative test examples are excluded.
3. Predictive Accuracy = (TP+TN/ Total)*100%= (TP+TN/TP+TN+FP+FN)*100% i.e. a high level of confidence can be placed only for results that give high values for all three measures.
4. Precision= TP/ (TP+FP)*100%, how many of the test correctly classified from the total test and Recall= TP/(TP+FN)*100%, how many of the actual correct value classifies correctly.

CHAPTER FOUR

BUSSINESS AND DATA UNDERSTANDING

One of the phases in the knowledge discovery process is understanding the business domain. Without a keen understanding of the business domain, no matter what tools used or how good techniques followed, may not provide useful result(15). Having an in-depth knowledge in the business domain enables data analysts clearly set the objectives and attempts to be made to attain the defined goals. This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives(66).

4.1 Blood Donation Process

Donating blood is safe and simple. The entire process, from registration to refreshments takes approximately 30 minutes. Strict procedures for every step of the process ensure the safety of the donor and of the blood supply. Every measure is taken so that the donation is safe for the donor and the blood recipient(17).

In general the overall processes followed in the course of blood donation are presented below(17):

Step 1: Registration: In the reception area, general information (name, address, age, sex…..) is recorded. In order to maintain accurate records, all donors are asked to present their proper identification.

Step 2: Medical interview: Every donor meets privately with blood bank staff members to review his/her medical history and this information is kept confidentially.

Step 3: Mini physical check: During the mini-physical check the weights and pressures of individuals are usually checked to assure if they fit the minimum requirement and a drop of iron

will be taken from the finger and tested to make sure there is enough iron-carrying blood cells to safely donate blood.

Step 4: Every donor is taken to the actual donation area, where a phlebotomist sterilizes the area of donors' arm from which the blood is drawn.

Step 5: After donation the donor is directed to the canteen area, where he/she can take a rest for approximately 15 minutes and is allowed after wards to leave and resume his/her daily routine.

Step 6: Having collected every unit of blood under goes very intensive screening for measuring the TTI's (HIV, Syphilis, Hepatitis B, Hepatitis C screening).

In the Data Analyzing Unit of the ENBBS the Donors data is usually on a monthly and yearly bases analyzed if the collected blood meets the target plan. By target plan it is meant there is a yearly plan to collect a certain minimum amount of blood from voluntary or family replacement donors' and the total collected blood is compared with the intended demand and accordingly will be reported to the WHO, FMOH, and the NBBS directorate(17).

Apart from the reporting system, simple queries are used to retrieve list of donors usually family replacement donors' to make sure if they have a history of donation and receive blood from the stock up on request.

It is evident that, there has been gross inadequacy and in-equitability in access to blood safety by the population, particularly in the regions. Only 24,000 units were collected in 2004 (i.e. 0.3 units/1000 people) and of these 17,000 units (71% of the total) were collected from Addis Ababa. This indicates the severe shortage of blood supplies for the vast majority of the population (about 96%) residing outside Addis Ababa(78).
Moreover, Screening for transfusion transmissible infections (TTI's) to exclude blood donations at risk of transmitting infection from donors to recipients is a critical part of the process of ensuring that transfusion is as safe as possible. Effective screening for evidence of the presence

of the most common and dangerous TTI's can reduce the risk of transmission to very low levels(18).

Currently one of the means to identify the eligibility of blood donors in fact the family replacement donors to donate a blood before the actual screening the blood for TTI's is a direct interview regarding donors personal behaviors and previous sign for any long duration diseases donors with noticeable signs and previous risky behaviors are excluded from the actual donation .However, with the very intricate nature of human beings and less interest to explain past experiences makes the donor recruitment process less efficient.

The very intensive screening criteria are taken before the actual donation is held are categorized in to the inclusion and exclusion criteria(17).

- The exclusion criteria is the mandatory requirement a donor ever asked before donating a blood and includes minimum weight 45, age above 18 and currently non pregnant for female donors.
- Having satisfied the exclusion criteria the second phase is the inclusion criteria to assess the eligibility of the donor before undergoing the donation. Under this phase oral interviews are maintained between the donor recruitment personnel and the donor if he/she is capable of donating.

Table 4.1 Registration datasets used in donor screening

Name	Used in order to keep name of the donor.
Sex	This is captured for statistical information and is not a criterion for donation as long as females are not pregnant and lactating mothers.
Age	The cutoff point for a donor age is 18 and 65 and individual donors in between (18≤x65,where x= age of the donor), are eligible to donate.
Address(region,city subcity,kebelle)	Attributes are used to store the address of the donor and all attributes having a separate column. Even though the National Blood Bank is in Addis Ababa, donors' might come from outside Addis Ababa and the exact regions have to be recorded explicitly using the attribute region.
Date	Attribute indicates the exact time when the donation was done. It has the format of dd/mm/yy.
Weight	Is used as exclusion criteria for blood donors the minimum cutoff point for a blood donor in order to be eligible to offer a blood is 45 Kg and above 45 Kg is possible.
Occupation	It explicitly defines the occupation of individuals.
Donation type	It helps to identify whether the donation is voluntary or family replacement.
Site of donation	It indicates the different blood donation sites such as schools. colleges' camps and associations.

The common questions asked includes: medical history of the donor such as respiratory system, cardiovascular system, rental system, central nervous system, the general health of the donor, weather the donor ever suffered from anemia, if the donor had ever been deferred from donating

and if there are noticeable medical signs of sexually transmitted diseases. All the questions are asked but not physical examinations are under way. All the questions asked are not a guarantee in maintaining a safe blood. Because naturally donors do not want to tell their experiences and there is a high possibility of bias from oral interviews in identifying candidate donor.

Table 4.2 Data captured in the course of donor interview

Disease type	Yes	No
Does the patient has any cardiovascular,rental ,and central nervous system failure?		
Is the patient diabetic or hypertensive?		
Is the donor pregnant or lactating (for female donors)?		
Is the patient epileptic?		
Does the donor has any chronic medical condition?		
Is the donor taking any prescribed drugs ?		
Does the donor have any widespread skin lesion?		
Does the donor have any history of donation		

In the process of blood donation the physical examinations are undergone after the donor has already donated blood, and this examination is aimed in maintaining a safe blood transfusion service not intended in maximizing the optima blood collection. The blood sample which has at least one of the transfusion transmission diseases is usually discarded.

Table 4.3 Blood donation ,testing process and datasets used

Date	The date attribute is used in order to determine the expiry date of the blood
HCV	Holds the HCV status of the donor after the physical examination and has parameters either "Y" or "N" indicating yes or no.
HBV	Holds the HBV status of the donor after the physical examination and has parameters either "Y" or "N" indicating yes or no.
HIV	Holds the HIV status of the donor after the physical examination and has parameters either "Y" or "N" indicating yes or no.
RH	Rhesus factor also known as Rh factor has Rh+ and Rh- values in the different blood types and it is used as a criteria in blood transfusion to determine compatibility issues.
ABO	Blood types such as "A","B" are recorded as they are important factors to determine compatibility issues.

The main reason that has necessitated this research is therefore, the data warehouse of donors data can be used to discover a hidden knowledge and patterns that can play a role in maximizing the safety of the collected blood. The data can be used not only for a reporting but given the demographic characteristics of donors; data analysis may uncover important data patterns, contributing greatly to business strategies in providing a novel knowledge that can be used as a base for guidance and decision making.

Predicting which age group, sex, blood type, type of blood donation (voluntary or family replacement) and location are more likely to have a safe blood would increase the likely hood of safe blood in stock and provide optimal blood on demand. Hence data mining technology can

offer immense potential in predicting and mining the hidden characteristics and patterns that exist in the blood bank such as the safest age group, sex occupation for the major TTI's. Different data mining classification algorithms and models can be trained and built a model to predict which age group, which blood type, which sex and which kind blood donation and which location are the safest for transfusion transmitted diseases. Classification of such patterns and knowing the safest group of the risky factors will enable the blood bank to empower the advocacy and consultation of blood donation services to the safe group which finally would maximize the collection of safe blood in the blood bank. Furthermore knowing the magnitude of the TTI's can help other stakeholders to guide their intervention and set policies accordingly.

4.2 Data Understanding

Domain experts were consulted to have a bird's eye view into the problem domain. The domain experts communicated includes two individuals from different departments namely from the Blood Donor Service Management Division and Data Analyzing Unit. The former department is concerned with the whole process of blood screening up to the distribution of the blood in to different hospitals and the consultation of this department has presented what exactly the business is and what kind of data are captured during blood donation. The second division which is concerned with the data management has the function of keeping track of every donor's record and report the statistical information regarding the collected blood to concerned bodies such as HAPCO (HIV/AIDS Prevention and Control Office), MOH (Ministry of Health),WHO (World Health Organization) and the directorate office of ENBBS(Ethiopian National Blood Bank Service) .

4.3 Data Source

The data employed in this research was collected from the Ethiopian National Blood Bank Service (ENBBS). A full backup of the database of the blood donors' database of the ENBBS was taken.

Initially, information about the blood donor is recorded when the individual arrives at the reception which includes the demographic characteristics of individuals and their prior medical history to assure their eligibility to donate and the information is recorded in information collection sheet. In fact, all information clerk personnel at the reception area are provided with a centralized form or record format that should be filled when a donor is to offer a blood and this helps to maintain consistent information.

After the blood donation, as explained in section 3.1 blood samples are taken from each donor's serum to screen for the major transfusion transmittable diseases such as HIV, HBV, HCV and syphilis. Donated blood with at least one of these diseases is stored in the table of infectious.

The blood donation database of the ENBBS contains more than 150,000 total records and 14735 with seropositive for at least one particular TTI's diseases.

But, still there are a number of records stored manually which needs to be captured to the automated system. Since large volume of data is more important to train data mining models(17), for the research also the researcher has taken the available records with seropositive amounting to 14757 and records accounting to 2107 without any seropositivity .The later figure that corresponds to the safe group blood donors record is taken in order to represent the safe groups sample in the model development.

4.4 Statistical Summary of Attributes

For data preprocessing to be successful, it is essential to have an overall picture of the data pertinent at hand. Descriptive data summarization techniques can be used to identify the typical properties of the data and highlight which data values are the predominant. Furthermore, it can underline the missing values, outliers what method to follow in replacing them. The sample record that is used to denote the safe blood group used for the analysis is without any missing

value and outliers. Hence the statistical summary is devoted for the records with at least one seropositivity.

Under the different attributes described in table 3.1 and 3.3., date is one of the variables used to keep the exact date, month and year of donation. The ENBBS started to automate their system and to store records in a computerized way since 1996.Therefore, the proportion of infectious diseases in each year starting from 1996 until 2003 is described in the table 4.4.

Table 4.4: Frequency of disease from 1996-2003

Year	Frequency of disease	Percent
1996	1618	10.99%
1997	2814	18.99%
1998	1993	13.5%
1999	2165	14.6%
2000	1890	12.7%
2001	1948	13.2%
2002	2064	14.0%
2003	265	1.82%

Age distribution of donors' under the different categories is presented in table 4.5

Table 4.5: Donors by age categories

Age	17-30	31-40	41-50	51-60	>60	Missing value	Outlier
Frequency	8724	4212	1436	352	28	4	1
Percent	59.1%	28.5%	9.7%	2.38%	0.189%	0.027%	0.0067%

Table 4.6: Summarizes the frequency of Sex distribution among the donors

Sex	Frequency	Percent
M	12181	82.5
F	2576	17.5

Frequency of occupation of donors by different categories are recorded as presented in table 4.7

Table 4.7: Donors by occupation

Occupation	Frequency	Percent
Civil Servant(Cs)	1974	13.37%
Private worker(Pw)	6530	44.2%
House wife(Hw)	391	2.64
Private employee(P.emp)	598	4.0%
Daily laborer (DL)	131	0.88%
Housemaid servant(Hm)	22	0.14%
Driver	402	2.72%
Farmer	335	2.27%
NGO	222	1.5%
Religious	46	0.31%
Student	3065	20.76%
Unemployed	789	5.34%
Missing value	4	0.02%

Table 4.8 shows the distribution of blood donors' location. Since it is a research conducted at the National Blood Bank of Ethiopia centered in AA, there is a possibility of donation from individuals who come from the different regions to AA. Therefore, the address indicates donors' from AA and other parts of the nation.

Table 4.8: Frequency of donors by region category

Address	Frequency	Percent
Fourteen(AA)	13476	91.3%
One(Tigray)	23	0.15%
Two(Afar)	13	.088%
Three(Amhara)	174	1.17%
Four(Oromia)	891	6.03%
Five(Somalia)	33	0.22%
Six(Benshangul Gumuz)	5	0.03%
Seven(SNNP)	109	0.73%
Nine(Gambella)	17	0.11%
Diredawa	9	0.06%

The donation of blood is largely in AA which amounts to 70 % of the total blood collected in the country(17). However people from regions usually come and donate in AA mainly as replacement for a family indeed in need of blood. Table 4.9 shows donation of blood in AA both from residents and from cities selected as a sample in the country.

Table 4.9: Frequency of donors by city category

City	Frequency	Percent
AA	13398	90.7%
Bishoftu	134	0.9%
Other Oromia	163	1.1%
Arusi	46	0.31%
Burayu	22	0.14%

Holeta	30	0.20%
Adama	45	0.30%
Metehara	28	0.19%
Harar	24	0.162%
Missing value	10	0.06%

Table 4.10 Presents the contribution of the different sub cities in Addis Ababa.

Table 4.10: The frequency of donors by sub cities

Woreda	Frequency	Percent
Kirkos	1614	10.99%
Gulele	1275	8.6%
Arada	1468	9.9%
Kolfe Keranio	981	6.6%
Bole	1763	11.9%
Yeka	1629	11.0%
Nifassilak Lafto	1405	9.5%
Lideta	803	3.38%
Addis Ketema	925	5.4%
Akaki Kality	1144	7.7%

Table 4.11 Shows the weight distribution among the donors taking 45 Kg as being a minimum cutoff point.

Table 4.11: Donors by weight categories

Weight	45-55	56-65	66-75	76-85	86-95	>95	Noisy value
Frequency	2481	5473	3899	2052	661	190	1

Percent	16.8%	37.0%	26.4%	13.9%	4.47%	1.28%	0.06%

The biological classification of blood type that the human being can posses is one of the four namely: A,B, AB or O. Table 4.12 indicates the composition of blood types where "O" as the most predominant and "AB" as the least predominant.

Table 4.12: Frequency of donors by blood type

Blood type	A	B	AB	O	Missing value
Frequency	4415	3522	831	5969	20
Percent	29.91%	23.8%	5.63%	40.44%	0.13%

Rh factor is also one of the attributes used at the ENBBS which a screening is usually undergone to determine the compatibility of blood during transfusion process ; if an Rh negative person receives a blood transfusion from an Rh positive person it can result in hemolytic and anemia. Table 3.13 below shows the Rh "+" and Rh"-" distribution.

Table 4.13: Rh Frequency of donors

Rh	Yes(+)	No(-)	Missing
Frequency	13778	976	3
Percent	93.3%	6.61%	0.02%

As described in the introduction section there is a shortage of blood in the nation and the pressing demand to have an abundant blood has necessitated the advocacy to be geared towards voluntary blood donors. Table 4.14 Shows the ratio of the donation type. The largest share is donated by the family replacement (denoted by "Rep") followed by mobile(Mob) blood collected from different location out of the blood transfusion site and "Vol ", used to denote the voluntary blood donors'.

Table 4.14: Frequency by blood donation

Donation type	Vol	Mob	Rep	Missing
Frequency	1415	1841	11499	2
Percent	9.5%	12.4%	77.9%	0.01%

Table 4.14 indicates contribution of the different sites; the blood transfusion site being the place where large volume of blood is being collected. As mobile blood collection sites' are large in number; samples are used to view the descriptive aspects. However, all blood donation sites were used in the analysis phase. Stratified samples were drawn from different categories in order to show the distribution in selected sites.

Table 4.15: Frequency of donation by site

Donation site	Frequency	Percent
BTS	12826	86.1%
World Vision Ethiopia	51	0.34%
Admas College	25	0.169%
Bishoftu Preparatory	52	0.35%

Wondirad high school	89	0.60%
AAU	15	0.10%
AA Muluwengel believers' church	75	0.50%
Youth association	70	0.47%

Knowing the transfusion transmittable diseases' magnitude and patterns with respect to the different attribute values used in the ENBBS data base is what interested the researcher to conduct the research. The statistical description of the three TTI's diseases is enumerated in table 4.16.

Table4.16: Frequency of infectious diseases by category

Cases	**Yes**	**No**
Hepatitis B	**8208**	**6548**
HIV	**3587**	**11170**
Hepatitis C	**3950**	**10807**

As indicated in the different tables the mean value of missing values is from 0.02 to 0.06 which is insignificant when compared with real world data warehouse which suffer from missing values, outliers, typos errors and inconsistent data. However, it is worth mentioning there were inconsistencies in the column Subcity in recording addresses of donors. All in all the record was not so much challenging as the attributes were quite manageable in number.

4.5 Data Preparation

The ENBBS donor database system is developed with mysql with a front end application of a visual basic 6.0.The database comprises information about donors before the actual screening for the major transfusion transmissible diseases is undergone. This information is recorded in a table "donor table" and information about donor with at least one of the TTI's is recorded in the

table"infecteous" ,safe blood which is discarded because of compatibility, expiry and other technical issues is recorded under the table "discard". Since the information was kept in different tables and to prepare the data for analysis it was important to denormalize or merge some of the tables. For this reason information about TTI's, which is stored in the infectious table and information about the donors found in donor table were exported in to Excel file which the later donors' general information is used to determine the magnitude of the prevalence of the diseases from the different attribute value perspectives.

The ENBBS blood donors database system which is employed for the purpose of this academic research, suffers from a number of limitations which includes missing values, outliers and encoding inconsistency in various attribute values. In fact, as stated by Witten and Frank(79) , one of the serious problems in building data mining models is limitations in the data itself. Thus, an optimal model could be constructed once a comprehensive, clean and automated data is well prepared.

One of the most important tasks in data mining is preparing the data in a way that is suitable for the specific data mining tool or software package. Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process(66). Thus, data cleaning has become a must in order to improve the quality of data so as to improve the accuracy and efficiency of the data mining models.
The purpose of data preprocessing in knowledge discovery process is to cleanse the data and to transform it into a form that is suitable to the subsequent steps. The data preprocessing includes a number of tasks and the common tasks are presented and discussed as follows.

4.6 Data Cleaning

The first stage of the data preprocessing was handling inconsistent and missing values. Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses. Incomplete data can occur for a number of reasons having incorrect attribute values, data collection instruments used may be faulty which may result in

noisy data, human or computer errors occurring at data entry, and missing the most important attribute values.

Under the ENBBS database the inconsistency was mainly due to typing errors and some of them were originally unrecorded.

Most of the inconsistent values that took the overall time of the data preprocessing time were the attributes woreda, city and kebelle. For example there are 25 possibilities to record a donor's address where his/her address is from Kirkos sub city. The possible attribute value for this exact scenarios includes: Cherkos, Cherekos, K.K,k,k,K.ketema,kirkos and the like. In all the attribute values of woreda,city and kebele the different possible values were finally rewritten in to the more appropriate name with the correction of spelling errors and keeping the consistency in naming. From the columns Subcity,woreda and kebelle, the researcher came to learn there was duplication of column values. For example if the Subcity is kirkos the woreda is also recorded as kirkos and kebele with a digit representation.In here it is not possible to determine the location of the kebele as belonging to a certain woreda since the naming of kebeles is uniform along with different woredas'. For this reason the column subcity was taken for the data mining analysis by replacing the woreda names by their corresponding subcities. Furthermore all attributes have lacked consistency and uniformity. for example under the attribute occupation" Cs" and "cs" are used to denote the civil servants, "Student" and "student" to denote student, all other such inconsistent values were rearranged in one uniform representation.

What followed next after handling the inconsistent values was handling the missing values. Theoretically, there are several ways suggested in the data mining literature to handle missing values such as calculating the average of continuous valued attributes and filling this mean value to missing attribute values (80). Nevertheless, in this study records with the exception of the attribute woreda and kebelle sought to have very minor missing value in the range between 0.02 and 0.06 and these missing values were replaced by the global constant"999".

4.7 Attribute Selection

Hypothetically, decision tree could figure out relevant attributes for classification automatically using the concept of information gain or entropy without manual efforts. However, in real world databases, it is common to find a number of irrelevant attributes (attributes that could not be helpful for some task of classification or other purpose) that could be easily known their irrelevance before adopting any complicated technique. Thus, it is important to exclude those attributes that are not important for analysis in order to simplify the task of decision tree. Moreover, as the irrelevant attributes might contain missing and noisy values, the exclusion of such irrelevant attributes can save the time needed to handle the missing values in the data preprocessing phase.

The ENBBS database contains 16 attributes which in fact are not as many attributes used in many databases. Since the attributes were all important , there was no need of employing feature selection for the most important attribute.

In general all the attributes used by the data management unit at the Ethiopian National Blood Bank except woreda and kebele merged in to subcity were preprocessed to be used for analysis by data mining models.

4.8 Summery of Original and Target Datasets

The blood donation dataset used under this study has attributes: Date, Age, Sex,Region, Subcity ,Occupation,ABO,Rh,Donation type,Site of donation,HCV,HBV,HIV. Table 4.17 presents the comparative summary of both the original and target datasets.

Table 4.17 summery of dataset

Summery	Original Dataset	Target Dataset
Number of Attributes	16	14

File Format	.xls	.xls	.CSV	.arff
File size	21.5MB	495KB	480KB	476KB
Total Number of Records	16864	16864		

CHAPTER FIVE

EXPERIMENT AND ANALYSIS OF CLASSIFICATION MODEL

The end result of the data preprocessing is a data that is suitable for any data mining algorithm. The choice of the techniques to be followed strongly depends up on a good understanding of the tasks to be conducted. As stated in the general objective the goal of this study is to build a predictive model to classify the seroprevalence of TTI's at the national blood bank service of Ethiopia. In order to predict the seroprevalence of the TTI's; the use of classification algorithm such as decision trees(J48), Bayes(Naïve Bayes) available in Weka 3.72 becomes evident.

For the experimentation purpose, three candidate scenarios were presented to be built by different models.

- The first scenario was to build the different classes, a model that predicts for HBV,HCV,HIV separately by using the same datasets.
- The second scenario was to build a model that predicts two class labels: a class label that has at least one infectious disease and a class label that is safe. And this model would predict the safe and unsafe group.
- The third scenario was to formulate a distinction in the unsafe group in to critical (where donors' having HIV exposure), and medium unsafe (donors having Hepatitis), and a third class label the safe group.

The first scenario attempts to predict three class labels of HBV,HCV, HIV by using three different models; which means a donor has to undergo three different test for the three TTI's when applied in to the real world application. Looking at other options that integrate the three class label in to one model was blatant and the two scenarios proposed.
The third model tries to predict the diseases in a degree of severity: critical, medium unsafe and safe group. The critical class label represents those who have HIV exposure to be more devastating if their blood is transfused to recipients than those who have Hepatitis.
The choice of these two scenarios demanded further justification from the domain experts and they confirmed as there is no disparity in blood transfusion if a donor has one of the TTI's. Because of this reason the second scenario that classifies donors in safe and unsafe group was used as experiment set up.

What followed next after the preparation of the experiment set up was to load the data preprocessed in .arff (attribute relation file format) in to Weka 3.72.

14757	2107	14757	13941

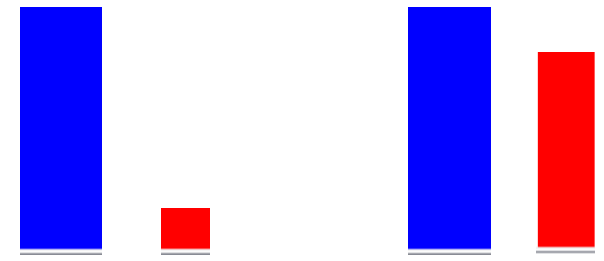

Figure 5.1 class imbalance problems and solutions

5.1 Issue of Class Imbalance Problems

As can be observed from the figure 5.1 the distribution of the dataset is skewed to the unsafe group making an imbalanced class.

A dataset is called imbalanced if at least one of the classes are represented by significantly less number of instances than the others(81). In imbalanced data classification, the class boundary learned by the standard machine learning algorithms can be severely skewed toward the positive class. Thus, the false-negative rate can be excessively high. One major task to overcome the class imbalance problem is to resample the original training dataset, either by oversampling the minority class and/or undersampling the majority classes until the classes are represented in a more balanced way.

Oversampling causes longer training time and inefficiency in terms of memory due to the increased number of training instances and it suffers from high computational costs for preprocessing the data. However, under sampling may discard useful data that could be important for the learning process (81).

Having analyzed the effect of oversampling and under sampling of the classes, the researcher has chosen the synthetic minority over sampling technique (SMOTE)(82). SMOTE as presented in section 3.2 is an important approach to resample the imbalanced class by over sampling the

positive class or the minority class to handle imbalance of class as shown in figure 5.1(after smote applied).

The oversampling causes the minority class to increases its number keeping the majority class constant. Subsequent re-sampling makes the minority class to approach the majority class or become a majority class making the former majority class to be a minor class. A perfectly balanced class is not attained because of the constant multiple of the minority class rather a compromised class where a significantly reduced class imbalance is achieved. From the figure 5.1 above the balanced class has reached the ratio of 94:100 (for 94 records in the minority class there are 100 records in the majority class) where the former imbalanced class was in the ratio of 14:100.

5.2 Building Classification Models

Four scenarios were intended to be built. These are binary decision tree with and without pruning and bayes naïve with display mode in old format as true and false. Trying those four scenarios with full attributes by changing the different parameters in the object editor of Weka 3.72 was performed.

These classification models are discussed in this section to compare their efficiency of the classifier models so that applicability of decision tree and bayes classification models can be discovered to the domain problem as stated in the objective section of chapter one.

In classification model building, validation (measuring classifier accuracy on unseen data) is very important issue. Weka provides three validation methods for model accuracy problems namely; cross validation, percentage split and the distinct training and test data sets.

The last two options have been used for this research. In the percentage split option, out of the total data prepared 70% (11805) records were employed for model building and the remaining 30% (5059) records were used for validation set or testing. Result of the test data indicated the model has 88% accuracy in predicting the unknown class labels.

5.3 Experimentation and Analysis of Results

Analysis of the decision tree and bayes models is made in terms detailed accuracy of the classifier on the training dataset as tested on the test data based on a confusion matrix of each model result. The confusion matrix is a valuable tool for analyzing how well our classifier can recognize tuples of different classes (True and False classes in the case of this research). Confusion matrix shows four important numerical quantities (true positive, true negative, false positive and false negative).

As it has already been discussed in Chapter 3 section 3.1.1 and 3.1.2, there are two scenarios to be experimented both the decision tree and bayes classification making a total four separate scenarios. These scenarios were experimented and analyzed to compare them to each other in terms of different performance matrices values, accuracies, number of leaves, size of tree generated, ROC curves and execution time. The models are also compared with regard to the patterns/ knowledge discovered. The scenarios for the aforementioned classifications and experimented in this research are as listed below: results of each model are analyzed and compared and finally selected the best model based on the criteria of evaluation.

Table 5.1: The different scenarios under taken to build the model

Scenario #1: Decision Tree with pruning
Scenario #2: Decision Tree without pruning
Scenario #3: Naïve bayes with DispalyModeInOldFormat False
Scenario #4: Naïve bayes with DispalyModeInOldFormat True

5.3.1 J48 experimental result analysis of occurance of unsafe blood donors

To predict the occurrence of unsafe group blood donors the object editor under Weka 3.72 provides the options of using MinNumObj(The minimum number of instances per leaf) with default value 2 and can be flexibly changed to increase the number of leaves under a given node and minimize successive tree branching. Furthermore, it also gives several options related to tree pruning.

In essence J48 employs two pruning methods. The first is known as subtree replacement. This means that nodes in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed subtree raising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way.

Error rates are used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential overfitting. This approach is known as reduced-error pruning. In order to assess the effects of MinNumObj and the confidence for pruning nine experiments were conducted.

Table 5.2: Values of parameters used in the nine Experiments

Experiments	**Parameters**		
	Pruned	**Confidence Factor**	**Numbers of Instance (minNumObj)**
Experiment #1	True	0.25	2
Experiment #2	True	0.25	5
Experiment #3	True	0.25	10
Experiment #4	True	0.30	2
Experiment #5	True	0.30	5

Experiment #6	True	0.30	10
Experiment #7	True	0.50	2
Experiment #8	True	0.50	5
Experiment #9	True	0.50	10

Performance	Experiments

As it can be observed from table 5.2 above the experimentation of J48 algorithm was conducted in nine different cases by changing the parameters. Each trial of the experimentation resulted in the generation different parameter matrix as presented in table 5.3 below.

Table 5.3 Experimental results of J48 Decision tree with different parameters

	#1	#2	#3	#4	#5	#6	#7	#8	#9
Accuracy (%)	88.5 %	87.6 %	86.1%	88.7%	87.9%	86.3%	**89%**	88%	86%
Mean absolute Error	0.166	0.18	0.20	0.16	0.17	0.19	**0.15**	0.169	0.19
Numbers of leaves	3074	2388	1738	3253	2533	1795	**4357**	3145	2319
Size of tree	3564	2733	1968	3773	2903	2032	**5048**	3598	2626
Time taken to build(sec)	0.34	0.29	0.26	0.53	0.3	0.26	**0.39**	0.47	0.29
AV. TP Rate	0.88	0.87	0.86	0.88	0.87	0.86	**0.89**	0.81	0.86
AV. FP Rate	0.12	0.13	0.14	0.11	0.128	0.144	**0.114**	0.124	0.14 1
AV. Precision	0.88	0.87	0.86	0.88	0.88	0.86	**0.89**	0.88	0.86
AV. ROC Area	0.92	0.92	0.92	0.926	0.922	0.91	**0.929**	0.926	0.92
AV. Recall	0.88	0.87	0.86	0.88	0.87	0.86	**0.89**	0.88	0.86

As can be seen from the table 5.3 the pruned j48 algorithm has generated relatively comparable model accuracies with varied parameters. Although unpruned was intended to be experimented, it is learned that it is not important at this exact scenario. This is because, though not recommended, unpruned is usually experimented if the pruned j48 experimentation results with small tree and leaf size that doesn't generate further rule in the form of if…then. But the experimentation above reavealed quite complex tree size and implies no further experimentation of unpruned J48. Accordingly, a thorough review of the experimented results indicates trial #7 with better model performance and is chosen for analysis.

Experiment #1: J48 pruned with confidence factor 0.25(default value)

Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves : 4357

Size of the tree : 5048

Time taken to build model: 0.7 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 25541 89.0551 %

Incorrectly Classified Instances 3157 10.9449 %

=== Detailed Accuracy By Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	**0.853**	**0.077**	**0.907**	**0.853**	**0.879**	**0.929**
3	**0.923**	**0.147**	**0.878**	**0.923**	**0.9**	**0.929**
Weighted Avg.	**0.891**	**0.114**	**0.891**	**0.891**	**0.89**	**0.929**

=== Confusion Matrix ===

a b <-- classified as

12593 2164 | a = 1

993 12948 | b = 3

The arguable guideline for calculating correctly classified instances and incorrectly classified instances, as it is described in the methodology section of 1.6 are the confusion matrix. It makes sense to observe the detailed accuracy measure of this scenario for calculating accuracy measures and performance as presented above:

The numbers of true positives in this confusion matrix are 12593 records. Those records which were predicted as 'True' class by the classifier and also happened true by when tested on the test data are (True Positives). The number of the records which were classified to the "False' class by the classifier and they are actually False as tested on the test data (True Negative Rate) are 12948. The sum of TPR (12593) and TNR (12948) gives us correctly classified. The total number of the records which were correctly classified to true and false classes of the TTI's are 25541.

One can analyze from Table 5.3, J48 Tree with pruning scenarios has generated model with better performance though its tree structure is a bit complex and can be visualized and comprehend hardly. It has more leaf nodes as well as it is long lengthy. However, its ability in correctly classifying records into both 'True' and 'False' classes is good (89.05%). Its ROC area is also above 0.5, which is the minimum possible acceptable value for ROC curve. If drawn the ROC Curve 0.929 is above the diagonal. ROC area is plotted from True positive Rate (TPR) on the y axis against the False Positive Rate (FPR) on the x-axis. In a case where the ROC is 0.50 it means that there is a 50-50 chance for TPR and FPR, which is quite unacceptable.
Based on these criteria it is possible to evaluate the performance of this model from the value of ROC area is above 0.5, which is 0.929 and we can say the model has a promising accuracy.

5.3.2 Naïve Bayes experimental result analysis of occurance of unsafe blood donors

The second model experimented for predicting the seroprevalence of the TTI's was the Naïve Bayes algorithm. The object editor under this algorithm has fundamental parameters such as DisplayModeInOldFormat.This parameter is used depending on the number of classes and attributes we have. The old format is better when there are many class values and the new format is ideal when there are fewer class and many attributes. Like the J48 algorithm the naïve bayes was experimented in two scenarios.

Experimentation of Naïve Bayes with DisplayModelInOldFormat False

Test mode: 10-fold cross-validation

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

Naive Bayes Classifier

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 18941 66.1373 %

Incorrectly Classified Instances 9757 33.8627 %

=== Detailed Accuracy By Class ===

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	0.573	0.262	0.657	0.573	0.613	0.731
3	0.738	0.427	0.664	0.738	0.699	0.731
Weighted Avg.	0.661	0.35	0.661	0.661	0.659	0.731

=== Confusion Matrix ===

a b <-- classified as

8463 6294 | a = 1

3463 10478 | b = 3

It is quite imperative to see the results generated by the Naïve Bayes by changing the parameters of the display mode in old and new formats. Except there is a difference in the time taken to build the model; both the formats have the same learning capacity for the datasets trained and tested. This necessitates comparing the result of J48 and Naïve Bayes.

5.3.3 Comparison of J48 and Naïve Bayes models

Comparison of the two models is made in terms of the general model accuracy, detailed accuracy by class such as the precision,ROC Area,recall and the rules generated for interpretation. The following table 5.4 gives the relative comparison between the two models.

Model	Accuracy	Number of leaves	Size of tree	Time taken to build	AV.TP.Rate	AV.FP.Rate	AV.Precision	AV.ROC.Arae	AV.Recall
Naïve Bayes	66%	-	-	.06	.66	.35	.65	.731	.66
J48	89%	4357	5048	.39sec	.89	.114	.89	.929	.89

Table 5.4 Model comparison of J48 and Naïve Bayes

Table 5.4 shows there is a relative better model prediction in the case of J48 in correctly identifying the dataset. The ROC Area for Naïve Bayes indicates 0.73 lower when compared with the ROC Area under J48 which accounts 0 .929.This signifies the number of correctly classified datasets are higher in the model built by J48 than the Naïve Bayes. The overall model accuracy of J48(89%) shows it has better prediction. The relative better performance of J48 algorithm can be attributed to the nature of the data such as the handled missing values; the data consistency etc.Naive Bayes has a better prediction if the attributes are conditionally independent to each other. For the given data under study J48 has shown better accuracy and the rules generated by this model are used for interpretation. It is worth mentioning however, Naïve bayes is also a candidate to be used for predicting the seroprevalence even though its performance is relatively low.

5.3.4 Expert and classifier judgments

One of the steps in the knowledge discovery tasks is to evaluate the performance of the system in terms of how correctly the model classifies records in to different labeled classes. Sometimes, the expert and classifier judgments may differ in predicting a record to a certain class label. This shows that the record that is labeled by expert to one class may be labeled by the classifier to other class. This kind of phenomena often reduces the performance of the system. Under this study J48 has an accuracy of 89% which means 11 % of the total records are incorrectly classified by the J48 classifier. The classifier predicts the records in to a certain class as there are similar attributes that lie in the same class boundary. But the attribute that determines the class boundary of the given record is suppressed because of the data-driven (attribute similarity) trend applied by the classifiers. It is such kind of attributes similarities that results in the incorrect classification of records. In the present study, the expert and classifier vary in classifying a certain records. For instance in the table below, the expert classified the blood donor as unsafe if the donation is in the year 1997 and safe if the donation is in the year 1996 with the same parameter value of the other attribute. Here what determines the class label is the year of donation. On the other hand, the fact that all the values of the attributes except the year are the same, then the classifier takes the similarity of all the attributes disregarding the difference in the year of donation and classifies both to be unsafe. Likewise in the other sample records presented in table 5.5, there is donation type, year variation in the second and the third records but the model predicts to the safe and unsafe group respectively. All in all the misclassification arises from the underestimating of a single attribute's value taking the similarity of the other attributes as the predominant predictive values.

Date	Age	Sex	Occupation	Region	Subcity	Weight	Abo	Rh	Donation Type	Site	status	
											expert	classifier
1997	1	1	16	14	2	2	4	1	1	2	1	1
1996	1	1	16	14	2	2	4	1	1	2	3	1

1999	2	1	13	14	8	2	4	1	2	1	1	3
1999	2	1	13	14	8	2	4	1	0	1	3	3
1998	1	0	16	14	8	2	4	1	2	1	1	1
1999	1	0	16	14	8	2	4	1	2	1	3	1

Table 5.5: Sample records that shows classifier and expert judgments variation

5.3.5 Rules generated by J48

What makes the J48 algorithm a choice of data mining practitioners is it provides rules in the form of if….then which are easy to comprehend and be acquainted with most important rules obtained. Under this study, results generated by J48 algorithm and that are interesting in the classification of the records are presented for discussion.

Rule 1 if Region = 14 and blood donation type = family replacement and age is between 25 to 34 and sex = male and Rh status is positive then it is probable to be unsafe (124.0/26.0)

Rule 2 if Region = 14 and blood donation type = family replacement and date of blood donation = 1999 and subcity = bole and donors are civil servants then it is probable to be unsafe (31.0/2.0)

Rule 3 if Region = 14 blood donation type = family replacement and date of blood donation = 1997 and subcity = kolfekeranio blood type = "O" and age is less 25: then it is probable to be unsafe (21.0)

Rule 4 if Region = 14 blood donation type = family replacement and date of blood donation = 1997 and Subcity = akakikality and donors are male civil servants and then it is probable to be unsafe:(10.0)

Rule 5 if Region = 14 and blood donation type = mobile and donors are students and Subcity = bole: then it is probable to be unsafe (110.0/1.0)

Rule 6 if Region = 14 and blood donation type = mobile and donors are students and subcity = addis ketema and age = 2: then it is probable to be unsafe (79.0)

Rule 7 if Region = 14 and blood donation type = family replacement and occupation=privately employed and date of blood donation =1998 and subcity = lideta it is probable to be safe (30.0)

Rule 8 if Region = 14 and blood donation type = mobile and occupation = students and site of donation are colleges then it is probable to be unsafe (169.0)

Rule 11 if Region = 14 and blood donation type = family replacement and occupation=drivers and blood type="A" and subcity = nifassilklafto: it is probable to be unsafe (33.0)

Rule 12 if Region = 14 and blood donation type = family replacement and occupation= farmer and Rh status is positive and subcity = arada and site = 2: it is probable to be unsafe (25.0)

Rule 13 if Region= 14 and date of blood donation = 1998 and site of donation = blood transfusion site and Rh status is positive and donation type = family replacement and blood type = "A" and occupation = business owners it is probable to be unsafe (97.0/27.0)

Rule 14 if Region= 14 and date of blood donation = 1998 and site of donation = blood transfusion site and Rh status is positive and donation type = family replacement and blood type = "A" and occupation = student and age between 25 to 34 it is probable to be unsafe (11.0/3.0)

From the above rules it is possible to learn rule #1,5,8,9,12,13,14 to be the most critical rules generating better knowledge for classification of the TTI's. Attributes such as occupation, site of donation, blood donation type, age, sought to be the most important variables in classifying the records in to the respective safe and unsafe class labels.

5.4 Discussions of Results on Transfusion Transmittable Infections

In order to reach a common plat form about the very significance of the above rules and the attributes used to create those rules, the associations of the attributes with the predicted class predicted by rules were evaluated based on suggestions offered by domain experts and results of previous research works.

As it can be seen from the rules, the model has generated class predictions for all the predefined classes of the safe and unsafe group. Therefore, the discussion is made in a way it addresses results of all classes mainly donors with TTI's.

Results of rules in each class indicated that the majority of the blood donation which accounts for 82.4 % is in blood transfusion site where donors as family replacement come to donate when their families are under stress strain and admitted to hospital. Thus, as the donors don't have knowledge about their current status of HBV,HCV and HIV, there is a high probability of donating blood even if it is unsafe. Moreover, the rules further signify that blood collected from mobile donors(donors out of blood transfusion site) are high probable to have the TTI's than voluntary donors. Domain experts agreed with this result, as voluntary donors do usually have check up for their status it is arguable that they are in less extent to have any of the TTI's.

Rules indicated that there are disparities in seropositivity of the infections from occupation point of view. It is possible to learn that business owners were said to be the most exposed group to the TTI's than others. Domain experts inclined to agree on this result; that is business owners have the luxury of financial freedom and with the chance of having more than one sexual partners. Furthermore, domain experts further stated that college students are becoming victims of sugar dads and firm owners; this could expose them to TTI's. Results of the study have also confirmed that college students are having more prevalence of TTI's next to male civil servants. Drivers and unemployed donors were also identified as having high susceptibility of the infections next to business owners, civil servants and college students.

Age groups between 24 to 35 were found to be the most vulnerable group for the TTI's. This might be due to the fact that their age makes them be the most sexually active individuals. The age groups from 35 to 45 were also the next age category to have high prevalence.

It is well known that the Rh negative individuals are rare and the majority of the blood donors who are Rh positive accounts for 76.5 %. Thus, it was possible to learn from the results that

generated Rh positive individuals have high prevalence. Furthermore, blood type “O” is regarded as the universal donor, and this has made the majority of the donors (46%) be “O” type. Results of the study have shown that “O” type blood is the most exposed group. The researcher believes that this could be because of the fact that there are more blood”O” donors than the other blood type. Domain experts were astonished with the results of blood type “A” as the next blood group (next to blood type “O”) that are having more prevalence of TTI’s. Thus domain experts firmly argued that there is no positive correlation with blood type and TTI’s and this may call for further research.

Unexpected rules such as seropositivity of religious personnel and people coming from abroad have absorbed both the researcher and the domain experts. It is found that donors whose occupation is in the religious circle had the exposure to the TTI’s. Domain experts suggested that the so-called religious individuals might have multi sexual partners. Apart from the role in the unsafe blood transmission the result might call social issues. Further more people coming from abroad and who have donated blood, were said to have risk of exposure to TTI’s. This might arise from unsafe sex exposure while they are inland or their prior exposure before coming home land.

Rules generating on regions were not important because of the fact the majority of the donation is in AA, and the rules generated are expected results. Some of the

From the total of 156729 consecutive blood donors, 14757 (9.41%) had serological evidence of infection with at least one pathogen and 29 (0.19%) had multiple infections. The overall seroprevalence of HIV, HBV and HCV was 2.29%, 5.23%, and 2.30% respectively .the prevalence of HIV in the national blood bank has shown gradual decrease when compared with similar study undertaken in Gondar referral hospital(with prevalence of 3.8) by Belay Tesema(18) and there is a slight increase in HBV (with prevalence of 5.23) and gradual increase in HCV (with prevalence of 0.7).

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

Conclusion

The discovery of transfusion transmissible infections (TTI's) has heralded a new era in blood transfusion practice worldwide with emphasis on fundamental objectives of safety and protection

of human life. Blood safety remains an issue of major concern in transfusion medicine in Ethiopia where national blood transfusion services and policies, appropriate infrastructure, trained personnel and financial resources are inadequate.

Screening for transfusion-transmissible infections (TTI's) to exclude blood donations at risk of transmitting infection from donors to recipients is a critical part of the process of ensuring that transfusion is as safe as possible. Blood transfusion services should therefore establish efficient systems to ensure that all donated blood is correctly screened for specific TTI's.

The objective of this study was to explore the immense applicability of data mining technology in the Ethiopian National Blood Bank Service by developing a predictive model that could help in the donor recruitment strategies by identifying donors that are at risk of TTI's which can help in the collection of safe blood group which in turn assists in maintaining optimal blood.

Under the course of the research the methodology employed was hybrid knowledge discovery process model (KDD). A total data set of 156729 was collected among which 14757 with at least one pathogen and 2107 safe group was extracted for analysis. The fact that , data warehouse suffers from inconsistency, missing value and etc. Before the data are used by any algorithm the data preprocessing activities such as handling missing value, feature selection and discritization were performed.

Predictive data mining technique was selected for classifying the data sets. Several models were built by implementing the J48 decision tree and Naïve Byes classifiers. Experimental result shows that J48 decision tree classifier using pruned technique with default confidence factor at 0.25 and minimum numbers of instance at 10 performed best with accuracy of 89%. Best attributes such as occupation,blood donation site,age,blood type found to be the most critical variables in classifying the tuples in to the safe and unsafe group.

Interesting rules such as if blood donation type = family replacement and age is between 25 to 34 and sex = male and Rh status is positive then the probability to be unsafe were observed.

The results obtained in this research work have proved the immense applicability of data mining technology in predicting the seroprevalence of TTI's. Important rules were generated for the vulnerability of the TTI's. Results generated from J48 classifier algorithm have revealed data mining technology can provide a huge potential in the donor recruitment strategy and can provide a base and guidance for policy makers.

Although, promising and appealing results were achieved by J48 classifier, some records were classified incorrectly due to the data driven classification trend of the algorithm. This makes the reliance on the algorithm to be questioned.

Recommendations

- J48 decision tree algorithm has produced promising results by generating easily comprehended rules. In order to enhance the performance of the predictive model the proper research need to undertaken by employing neural network and support vector hierarchical classifier.
- Syphilis virus becomes in active three days after donation of the blood, and analysis is made excluding syphilis. It is recommended to conduct study on the active serum of primary blood donors to measure the burden of syphilis virus.

- Blood donors data representing the regional branches are reported in a final analysis in the form of excel to the national blood bank service centered at Addis Ababa, and analysis of raw data can't be made. It would be important to have a centralized data management through distributed system so that raw data representing the national blood donor can be made available in the center and analysis would be representative of the national blood bank donors.

- The records under this study are taken from the operational database of the ENBBS. However, as substantial data are needed for data mining tasks , further researches that integrate the operational and non operational data warehouse would rather come up with interesting results.

- It is observed that, there are inconsistencies in recording donors' data caused by lack of data integrity constraint and resulting in the data preprocessing taking extended time. It is recommended that the database is redesigned taking in to account the problems pertinent at hand.

References

1.ANAGAW S. Application of data mining technology to predict child mortality patterns: the case of butajira rural health project (brhp) . Masters thesis Addiss Ababa University,Addis Ababa,Ethiopia,2002.

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

3.Prather J. Medical Data Mining : Knowledge Discovery in a clinical Data Warehouse. Available URL:<http://www.amia.org/pubs/symposia/D004394.PDF>;2001. [Acess date August 04,2010].

4.Raghavan VV, A Perspective on Data Mining. *Journal of the American Society for Information Science*;49(5): 397-4021998.

5.Fayyad U, Piatetsky-shapiro, G. and Smyth, Padharic. From Data Mining to Knowledge Discovery in Databases. [database on the Internet];1996.[Access date August 04,2010]

6.Larvac N. Data Mining in Medicine : Selected Techniques and Applications. Available URL.: <<http://citeseer.nj.nec.com/lavrac98data.pdf>>.;1998.

7.Bresnahan J. Data Mining in the Health care : A Delicate operation. Available URL : <http://www.Cio.com/archive/061597-mining-content.html>;1997.[Acess date August 04,2010].

8.OdmsibbisIoSa BaM. Data Mining Applications in Clinical Outcome Measurment: CE, Department of Electrical and Electronics Engineering F, University of Macau, Taipa, Macau Available online at www.sciencedirect.com [Acess date August 05,2010]

9.Butch SH. Computerization in the transfusion service. VoxSanguinis, . 83(suppl 1), 105-110(2002).

10.Sundaram TSaS. Application of CART Algorithm in Blood Donors Classification PG and Research Department of Computer Science, DG Vaishnav College,Chennai-600106, Tamil Nadu, India.Available at: www.ijcset.net/docs/Volumes/volume1issue1/ijcset2011010103.pdf

11.Butch SH. Computerization in the transfusion service. Vox Sanguinis. , 83(suppl 1), 105-110(2002).

12.Butch, S. H. (2002). Computerization in the transfusion service. VoxSanguinis, 83(suppl. 1), 105–110.

13.Rea A. Data Mining An introduction Student Notes. <<http :// www pccqub acuk /tec / courses/datamining/stu_notes/dm_book_1html>>;2002.[Acess] date febraury 10,2011].

14.Levin NaZ, Jacob,. Data Mining. Available URL:www.urbanscience.com/Data Mining.pdf;1999.

15.Helen T. application of data mining technology to identify significant patterns in census or survey data. Masters Thesis Addis Ababa University, Addis Ababa,Ethiopia; (2003).

16.Dhingra N. Screening Donated Blood for Transfusion- Transmissible Infections:World Health Organization available at: << http://www. who.int/bloodsafety/ makingsafe bloodavailable in africastatement.pdf>>;,2002.

17.The Ethiopian Red Cross Society,National Blood Bank Service Highlights Blood a Gift for Life.. (2010).

18.Belay T. Seroprevalence of HIV, HBV, HCV and syphilis infections among blood donors at Gondar University Teaching Hospital, Northwest Ethiopia: declining trends over a period of five years;(2010).

19.Baye GaY, Mengistu,. The prevalence of HBV, HCV and malaria parasites among blood donors in Amhara and Tigray regional states.pdf;2002.

20.Dons SMaW, Michael W. Mining Association Rules from a pediatric primary care Decision support system. Available URL:www.amia.org/pubs/symposia/D200658.PDF <http: //www. amia.org/pubs/symposia/D200658.PDF> ;2000.

21.Scientific Methodology Definitions available at: <http://irn.uit .tufts.edu/research _planner /documents/6/methodology_tips.pdf>.

22.C, K.g;Pedrycz,W.j;Swiniarski,Rw;Kurgan,. Data Mining a knowledge discovery approach.pdf; 2007.

23.Two Crows Corporation Introduction to Data Mining and Knowledge Discovery.3rd ed Available URL:http://www.twocrows.com.[Acess date February,2011];1999.

24.Skalk D. Data Mining Blunders Exposed; Database programming and design Magazine. Available URL:<<http://www.db2mag.com/db_area/archives/2001/q2/miner.pdf>>;2001.

25.Tobler L BM. History of post transfusion hepatitis. Clinical chemistry.. 1997;43:1487-1493.

26.Medical Definition of Blood Available at:<<http://medical-dictionary.the free dictionary .com/Blood+preservation>>[Acess date February,10,2011]

27.Ethiopian Red Cross\Ethiopian Red Cross Society.mht\[Accessed date september 12,2011.

28.Ethiopian Red Cross\Ethiopian Red Cross Society.mht\accessed date [Acess date february 23,2011].

29.Tagny CT MD, Tapko JB, Lefrère JJ. Blood safety in Sub-Saharan Africa: a multi-factorial problem. . *Transfusion* 2008;48(6):1256-1261.

30.World Health Assembly resolution *Utilization and supply of human blood and blood products*. Geneva, World Health Organization, WHA2872:1975.

31.Koistinen J. Safe blood: the WHO sets out its principles. *AIDS Anal Afr*. 1992;2(6):4–6.

32.J. K. Safe blood: the WHO sets out its principles. *AIDS Anal Afr*. . ;2(6):4-61992.

33.Mbanya DN TD, Ndumbe PM. Serological findings amongst first-time blood donors in Yaoundé, Cameroon: is safe donation a reality or a myth? *Transfus Med*.. 13(5):267-273;2003;.

34.Bates I HO. Should we neglect or nurture replacement blood donors in Sub-Saharan Africa? *Biologicals*. . ;38(1):65-672010.

35.World Health Assembly resolution *Proposal to establish World Blood Donor Day*. Geneva, World Health Organization,. WHA5813:2005.

36.Bates I MG, Medina Lara A. Reducing replacement donors in Sub-Saharan Africa: challenges and affordability. *Transfus Med*;17(6):434-4422007.

37.Ejele OA NC, Erhabor O. . Seroprevalence of HIV infection among blood donors in Port Harcourt, Nigeria. . *Niger J Med*; 14(3):287-2892005.

38.Mumo J VA, Jehuda-Cohen T. Detecting seronegative-early HIV infections among adult versus student Kenyan blood donors, by using Stimmunology.*Exp Biol Med (Maywood)* ;234(8):931-9392009.

39.M. C. Emerging infections agents: Do they pose a risk to the safety of transfused blood and blood products? CDC, National center for infections disease, division of viral and Reckettsial disease. . 797-8052002.

40.*Aide-mémoire: Blood safety*. Geneva, World Health Organization, 2002.

41.*Consensus statement on screening blood donations for infectious agents through blood transfusion.* World Health Organization Global Programme on AIDS/League of Red Cross and Red Crescent Societies, WHO/LBS/9111991.

42.*Blood Safety Indicators*, Geneva, World Health Organization, . 2009.

43.Deogan,Data Mining: research Trends, Challenges, and Applications [database on the Internet]. <http://citeseer.nj.nec.com/deogun97data.html> [Accessed on february,21,2011].

44.Data Mining: research Trends, Challenges, and Applications [database on the Internet]. <<http://citeseer.nj.nec.com/deogun97data.html>> [Accessed date february 21,2011.].

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

46.Berry MaL, G. . *Data mining techniques: For marketing, sales and customer support*. New York. John Wiley and Sons, Inc. 1997.

47.Piatetsky-Shapiro G. Knowledge Discovery in Databases: 10 Years After. SIGKDD Explorations. Online. Internet.<<http://www.kdnuggets.com/gpspubs/sigkdd-explorations-kdd-10-years.html>>[access date february 30,2011]. 2000.

48.Piatetsky-Shapiro, Gregory. (2000). Knowledge Discovery in Databases: 10 Years After. SIGKDDExplorations.Online.Internet.<http://www.kdnuggets.com/gpspubs/sigkdd-explorations-kdd-10-years.html>.[Access date February,14,2011]

49.Last, Mark, Maimon, oded, and Kandel Abraham. 2002. Knowledge Discovery in MortalityRecords:Aninfo-fuzzyApproach.AvailableURL:<http://www. csee.usf.edu/ softec/med_dm3.pdf>.

50.World Health Assembly resolution WHA58.13: *Proposal to establish World Blood Donor Day*. Geneva, World Health Organization, 2005.

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

52.Levin, Nissan and Zahavi, Jacob, 1999. Data Mining:DATA MINING NOTES wwwurbansciencecom/Data_Miningpdf.

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

54.Tesfaye, Hintsay. (2002). *Predictive Modeling Using Data Mining Techniques In Support to Insurance Risk Assessment*. Masters Thesis Addis Ababa University, Addis Ababa,Ethiopia; (2002).

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

56.Thearling K. An introduction to data mining.*Available at:* <<http:// www3. shore. net/~kht/text/dmwhite.pdf>>;2003.

57.Chapman P.CRISP-DM 1.0 Step-by-step data mining guide SPSS Inc., U.S.A CRISPWP-08001999.

58.Data Mining a knowledge discovery approach, Cios,K .g;Pedrycz,W. j;Swiniarski, Rw; Kurgan,LA(2007).

59.Deogan. Data Mining: research Trends, Challenges, and Applications [database on the Internet]. <http://citeseer.nj.nec.com/deogun97data.html>;2001..[Access date January 6,2011]

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

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

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

63.Frohlich J. Neural Net Overview. Available at:<http://rfhs8012.fh-regenburg.de/~ saj39122/ jfroehl/diplom/e-text.pdf>;1999.

64.Arzucan ¨Ozg¨ur. Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization .Turkish: Bo˘gazin University;2004.

65.Feyen HaLP. Data Mining and Strategic Marketing in the Airline Industry. Online.[Access date January 10,2011].

66.Witten IHaF, Eibe. Practical Machine Learning Tools and Techniques with Java Implementations.USA: Academic Press. 2000.

67.WHO,(2007),- Regional training workshop on blood donor recruitment: pre and post Donation counseling Available at: <http:// www.who.i nt/countries /eth/ news/ 2008/ blood_ donor recruitment/en/index.html> [Access date October 08,2010]

68.Fayyad, Usma, Piatetsky-shapiro, G. and Smyth, Padharic. 1996. From Data Mining to Knowledge Discovery in Databases. [database on the Internet].[Access date February 08,2011].

69.Feyen, Hans and Lisa Pritscher (n.d.). Data Mining and Strategic Marketing in the Airline Industry. Online.[Access date February 04,2011]

70.Lloyd - Williams M. Discovering the Hidden secrets in your Data -the data Mining approach to Information. Available URL:<<http://informationr.net/ir/3-2/paper36.pdf>>; 1997.

71.Abraham T. Application of Data Mining technology to identify determinant risk factors of HIV infection and to find their association rules:the case of Center for Disease Control and Prevention(CDC). Unpublished Masters Thesis Addis Ababa University, Addis Ababa;2005.

72.University of Waikato"Weka Manual Version 3.6". 2008.

73.Quinlan JR. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA; 1993

74.Duad M.N.R aC DW. "Human Readable Rule Induction In Medical Data Mining: A Survey of Existing Algorithms",. WSEAS Eurpean Computing Conference,Athens,Greece 2007.

75.Ng AYJ, M. I. . On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, Neural Information Processing Systems, Ng,A.Y., and Jordan, M. 2002.

76.Viktor. HGaHL. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. SIGKDD Explorations,. 6(1):30{39, 2004.

77.Kowalczyk. BRaA. Extreme rebalancing for svms: a case study. SIGKDD Explorations, . 6(1):60{69,2004.

78.The Ethiopian Red Cross Society,National Blood Bank Service Highlights(2010),Blood a Gift for Life.

79.Witten, Ian H. and Frank, Eibe. 2000. Practical Machine Learning Tools and Techniques with Java Implementations. USA: Academic Press.

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

81.Japkowicz N. The Class Imbalance Problem: Significance and Strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning* Las Vegas, Nevada(2000). .

82.Bordes SE JW, and L. Bottou. . Fast kernel classifiers with online and active learning. Journal of Machine Learning Research (JMLR). . 6:1579-16192005.

Appendix

Rule 1 Region = 14 and Donation Type = 1 and and Occupation = 1 and Site = 4 and Subcity = 1: safe (169.0)

Rule 2 Region = 4: unsafe (1184.0) ,Region = 7: unsafe (132.0),Region = 3: unsafe (203.0)

Rule 3 Region = 14 and Donation Type = 1 and Occupation = 1 and Site = 4 and Subcity = 6: safe (80.0)

Rule 4 Region = 14 and Donation Type = 1 and Occupation = 9: unsafe (16.0)

Rule 6 Region = 14 and Donation Type = 1 and occupation=10 and date=1998 and Subcity = 7: safe (79.0)

Rule 7 Region = 14 and Donation Type = 1 and occupation=10 and date=2001: safe (162.0/4.0)

Rule 8 Region = 14 and Donation Type = 1 and occupation=12 unsafe (48.0)

Rule 9 Region = 14 and Donation Type = 1 and occupation=13 and subcity=2: unsafe (27.0),Subcity=3: unsafe (40.0)

Rule 11 Region = 14 and Donation Type = 1 and occupation=3 and ABO= 1 and Subcity = 9: unsafe (93.0)

Rule 12 Region = 14 and Donation Type = 1 and occupation=5 and Rh=1 and Subcity = 1and Site = 2: unsafe (95.0)

Rule 10 Region = 14 and Donation Type = 1 and occupation=13 and subcity=5 and Date = 2002: 3 (76.0/1.0)

Rule 11 Region = 14 and Donation Type = 1 and occupation=3 and ABO= 1 and Subcity = 9: unsafe (93.0)

Rule 12 Region = 14 and Donation Type = 1 and occupation=5 and Rh=1 and Subcity = 1and Site = 2: unsafe (95.0)

Rule 13 Region = 14 and Donation Type = 1 and occupation=16 and Subcity = 1and Age = 1 Date = 2002: 3 (391.0/11.0)

Rule 14 Region = 14 and Donation Type = 1 and occupation=16 and Subcity = 1and Age = 2: unsafe (15.0/2.0)

Rule 15 Region = 14 and Donation Type = 1 and occupation=16 and Subcity = 4: unsafe (110.0/1.0)

Rule 16 Region = 14 and Donation Type = 1 and occupation=16 and Subcity = 5 and Date = 1999: 1 (38.0) or Date = 2001: unsafe (34.0)

Rule 17 Region = 14 and Donation Type = 1 and occupation=16 and Subcity = 8: unsafe (185.0)

or Subcity = 9: unsafe (88.0)

Rule 18 Region = 14 and Donation Type = 1 and occupation=16 and Subcity = 10 and Date = 1996: unsafe (58.0) or Date = 1997: unsafe (83.0) or Date = 1998: unsafe (23.0)

Rule 19 Region = 14 Donation Type = 2 and Date = 1996 and Subcity = 3 and Age = 2 and Abo = 4 and Rh = 1: unsafe (20.0/7.0)

Rule 20 Region = 14 Donation Type = 2 and Date = 1996 and Subcity = 8: 1 unsafe (196.0/6.0)

Rule 21 Region = 14 Donation Type = 2 and Date = 1996 and Subcity = 9 and Age = 1: unsafe (22.0) and Age = unsafe :(55.0/2.0)

Rule 22 Region = 14 Donation Type = 2 and Date = 1996 and Subcity = 10: unsafe (121.0/3.0)

Rule 23 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 1: unsafe (183.0/14.0)

Rule 24 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 1 and Abo = 1: unsafe (42.0/1.0)

Rule 25 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 1 and Abo = 4: unsafe (34.0)

Rule 25 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and occupation=1 and sex=1: unsafe (39.0)

Rule 26 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and occupation=1 and sex=1 and occupation=18: unsafe (10.0)

Rule 27 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and occupation=1 Age = 3: unsafe (16.0/3.0)

Rule 28 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and occupation=1 and Abo = 4: unsafe (18.0)

Rule 29 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and Occupation = 3 and Abo = 1: unsafe (11.0/2.0)

Rule 30 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and Occupation = 5: unsafe (13.0)

Rule 31 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 3 and Occupation = 16: 1 (13.0)

and Abo = 4: unsafe (43.0/8.0)

Rule 32 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 6 Abo = 4 Age = 1: unsafe (21.0)

Rule 33 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 6 Occupation = 3: unsafe (16.0/3.0)

Rule 34 Region = 14 Donation Type = 2 and Date = 1997 and Subcity = 8 Occupation = 1: unsafe (33.0/2.0)

Rule 35 Region = 14 Donation Type = 2 and Date = 1997 and Occupation = 16: 1 unsafe (26.0)

Rule 36 Region = 14 Donation Type = 2 and Date = 1999and Subcity = 4 and Occupation = 1: 1 unsafe (31.0/2.0)

Rule 37 Region = 14 Donation Type = 2 and Date = 2000 and Occupation = 16: unsafe (10.0/2.0)

Rule 38 Region = 14 and Donation Type = 2 and Age = 2 and Sex = 1: and Rh=1 :unsafe (124.0/26.0)

Rule 39 Region= 14 and Date = 1998 and Site = 1 and Rh = 1 and Donation Type = 2 and Abo = 1 and Occupation = 16 and Age = 2: Yes (11.0/3.0)

Rule 40. Region = 14 Date = 1998 and Site = 1 and Rh = 1 and Donation Type = 2 and Abo = 1 and Occupation = 13: Yes (97.0/27.0)