American Journal of Artificial Intelligence
2017; 1(1): 44-55
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20170101.16

Predicting the Seroprevalence of HBV, HCV, and HIV Based on National Blood of Addis Ababa Ethiopia Using Data Mining Technologys

Haftom Gebregziabher[1], Million Meshasha[2], Patrick Cerna[1]

[1]Department of Information Technology, Federal TVET Institute, Addis, Ethiopia

[2]Department of Information Science, Addis Ababa University, Addis, Ethiopia

Email address:

habtilo@gmail.com (H. Gebregziabher), millionmeshesha20174@gmail.com (M. Meshasha), pcerna@acm.org (P. Cerna)

To cite this article:

Haftom Gebregziabher, Million Meshasha, Patrick Cerna. Predicting the Seroprevalence of HBV, HCV, and HIV Based on National Blood of Addis Ababa Ethiopia Using Data Mining Technology. *American Journal of Artificial Intelligence*. Vol. 1, No. 1, 2017, pp. 44-55.
doi: 10.11648/j.ajai.20170101.16

Received: May 14, 2017; **Accepted**: June 1, 2017; **Published**: August 30, 2017

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 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 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 TTIs 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 Naive bayes algorithm implemented in Weka. J48 decision tree algorithm with the overall model accuracy of 94% 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 TTIs was significant in business owners, students, civil servants, unemployed individuals, drivers and age groups 25 to 34 and 35 to 44 years.

Keywords: Data Mining, Blood Bank, HIV, HBC, HVC, CRISP-DM, Ethiopia

1. Introduction

1.1. Background

These days, people 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 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]. According to Bigus [2], 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.

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 system [3]. 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 bank [3].

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.

1.2. Statement of the Problem

The two crucial issues related to blood transfusion in the developing world, particularly Africa, are blood shortages and unsafe blood [7], 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 donation rates in Africa are generally very low (about 5 per 1000 population) compared with developed countries (for example, 47 per 1000 population in the United States). In its most recent global survey on blood safety and availability, WHO collected data from 40 of the 48 countries in sub-Saharan Africa [7]. This survey indicated that 35 (87.5%) countries collect less than half of the blood needed to meet the transfusion requirements of their populations. In 2004, only about 2.8 million units of blood were collected for a population of around 720 million people (11% of the world's population) [7].

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. Testing of the blood for the presence of major infectious pathogens such as Hepatitis B was not universal in most of the transfusion centers in the country. Although HIV was said to be tested in all of the units, problems in the supply of HIV testing kits was observed. There is also shortage of manpower in most of the centers [8].

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 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 in the major cities (such as Mekelle, Bahridar, Assosa) of the nation and this results the collection of safe and optimal blood inadequate [8].

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 [8].

Not less impressive 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 anchored towards the donors who are safe from infections.

Research conducted by Santhanam and shyam [9] 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 [10]: it is observed that a declining trends over a period of five years and the prevalence of HBV, HCV and malaria parasites among blood donors in Amhara and Tigray regional states [11] 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, among others. However, due to the limitation of statistical analysis such as their less 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 prevalence 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. Objective of the Study

1.3.1. General Objective

The main objective of this study is predicting the sero-prevalence, 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 Objective

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

a) Assess different classification, clustering and association rules mining application algorithms.
b) Select and extract the data set required for analysis from the Ethiopian National Blood Bank.
c) Preprocess: preprocessing data in order to have a cleaned dataset that is suitable for any data mining algorithm.
d) Train and build data mining models that help for predicting the sero-prevalence of HIV, HBV, HCV and syphilis at the blood bank.
e) Evaluate the performance of the data mining model using test data set and report findings.

2. Literature Review

2.1. National Blood Bank of Ethiopia

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 programs, the Federation aims to enable communities to reduce their vulnerability to disease, and prepare for and respond to public health crises [8].

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 countrysince 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, Bahir Dar, Gondar, dessie and Mekelle [8].

Currently the ERCS-NBBS has the following organizational structures dealing with specific activities [8]: Blood donor service management, laboratory division, quality control division, data analysis unit and administrative and finance unit:

a) Blood Donor Service Management Division is responsible for administering the donation and recruitment of blood to distributing blood and blood components to different hospitals.
b) Laboratory Division has a mandate of undertaking the screening of blood from transfusion transmittable infectious diseases.
c) 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.
d) 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

According to Tagnyet. Al [12] 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.

A study conducted by Baye Gelaw [11] 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.

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 [13]. Concerted efforts are still required by a substantial number of countries to achieve 100% screening of donated blood for TTI's within quality systems.

2.2. Data Mining Definition

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 from the databases information that might be of valuable to the owner of the database. The discipline concerned with the task has become known as Data Mining [14].

It has been estimated that the amount of information in the world doubles every 20 months [14]. 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 [14]. 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 [14].

2.3. 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. 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 [15].

KDD is; the nontrivial process of identifying valid, novel, implicit, potentially useful, and ultimately understandable patterns in data [15]. 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 [16] 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 [17], 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-Shapiro [15], 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 [18].

By grounds of the popularity of the term 'data-mining' than the term 'Knowledge Discovery in Databases', Han and Kamber [16] 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 [16] 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'.

2.4. 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 [19].

As Tesfaye [20] 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 [21]. 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 [22].

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 [22] 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 [21] 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 [21], 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 [21].

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 [21].

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 It was developed based on the CRISP-DM model by adopting it to academic research [21].

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. Many researchers and the area professionals have summarized this iterative process. Most of them agree that knowledge discovery process starts with a clear definition of the business problem or, equivalently, understanding of the application domain [23].

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 [23]. 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 [23]. The detailed explanation of the phases followed in each model is presented below. Figure 1 shows descriptions of the six steps of the CRISP-DM process model [23].

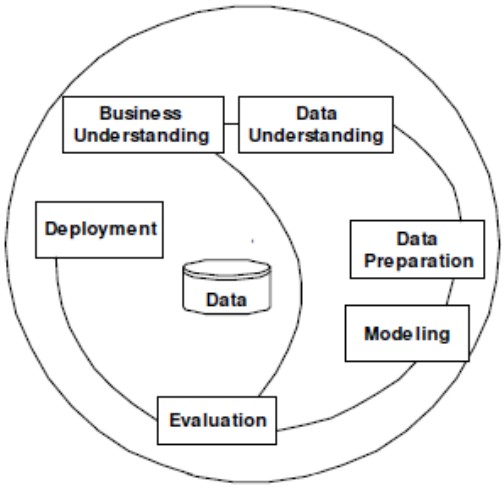

Figure 1. *The CRISP-DM process model.*

2.4.1. 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.

2.4.2. 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.

2.4.3. 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.

2.4.4. 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.

2.4.5. 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 [24], 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. The most commonly used data mining techniques are: Decision tree, neural networks, genetic algorithms, nearest neighbor method and rule induction [22].

According to Levin and Zahavi [25] 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 Han and Kamber [16] 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 [25]. As Han and Kamber [26] 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.

2.6. Application of Data Mining in Blood Dataset

Research conducted by Santhanam and Shyam [9] on application of CART algorithm in Blood Donors Classificationargued that the availability of blood in blood banks is a critical and important aspect in a healthcare system. Blood banks (in the developing countries context) 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

researchersidentified 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 [9].

A research was conducted by Baye Gelaw and Yohans Mengistu [11] 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.

3. Business 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 [19]. 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 [26].

3.1. 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 [8].

In general the overall processes followed in the course of blood donation are presented below [8]:

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).

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 section 1.3.1 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 (Naive Bayes) available in Weka 3.72 becomes evident.

Table 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.

3.2. Data Understanding

Domain experts were consulted to have a bird's eye view into the problem domain. The domain experts ever 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 Control Office), MOH (Ministry of Health), WHO (World Health Organization) and the directorate office of ENBBS (Ethiopian National Blood Bank Service). The data management unit division for data analysis purpose used the different variable shown in Table 1.

3.3. 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 is arrived 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.

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 [8], 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

3.4. 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.

Table 2. *Donors' by occupation.*

Occupation	Frequency	Percentage
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%

Frequencies of occupation of donors by different categories are recorded as presented in Table 2. Table 3 shows the distribution of blood donor's 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 3. *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 blood donation dataset used under this study has attributes: Date, Age, Sex, Region, Subcity, Occupation, ABO, Rh, Donationtype, Site of donation, HCV, HBV, HIV. Table 4 presents the comparative summary of both the original and target datasets.

Table 4. *Summary of dataset.*

Summary	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		

4. Experimentation and Analysis

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).

4.1. J48 Expermental 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 sub-tree 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 subtreeraising. 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. *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

As can be seen from the Table 5 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.

4.2. Naive Bayes Expermental Result Analysis of Occurance of Unsafe Blood Donors'

The second model experimented for predicting the seroprevalence of the TTI's was the Naive Bayes algorithm. The object editor under this algorithm has fundamental parameters such as Display Mode in old Format. This parameter is used depending on the number of classes and attributes we have. The old fprmat 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 Naive bayes was experimented in two scenarios.

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 6 gives the relative comparison between the two models.

Table 7 shows there is a relative better model prediction in the case of J48 in correctly identifying the dataset. The ROC Area for Naive bayes indicates 0.73 lower when compared with the ROC Area under J48 which accounts 0.92. This signifies the number of correctly classified datasets are higher in the model built by J48 than the Naive 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, Naive bayes is also a candidate to be used for predicting the seroprevalence even though its performance is relatively low.

a. 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)
b. 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)
c. 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)
d. 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)
e. 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)
f. Rule 6 if Region = 14 and blood donation type = mobile and donors are students and subcity = 1 and age = 2: then it is probable to be unsafe (79.0)
g. Rule 7 if Region = 14 and blood donation type = family replacement and occupation=privately employed and date of blood donation =1998 and subcity = 7: safe
h. 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)
i. 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)
j. 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)
k. 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)
l. 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)

4.3. Summary of Findings on TTI's

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 association 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 transfusion-transmissible infections (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 bussinesowners, 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 account 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 are 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") who 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.

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).

Table 6. *Model comparison of J48 and Naive Bayes.*

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
Naive Bayes	66%	-	-	.06	.66	.35	.65	.731	.66
J48	88%	3077	3564	.3 sec	.88	.12	.88	.92	.88

Table 7. *Experimental results of J48 Decision tree with different parameters.*

Performance Measure	Experiments								
	#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	0.141
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

5. Conclusion and Recommendation

The discovery of 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.

Predictive data mining technique was selected for classifying the data sets. Several models were built by implementing the J48 decision tree and Naive Byes classifier algorithm. Interchanging the parameters of the J48 and Naive Bayes classifiers generated different experiment scenarios.

All in all best performance was achieved by J48 decision tree classifiers using pruned technique, with default confidence factor at 0.25, minimum numbers of instance (MinNumObj) at 10 and with over all model accuracy 89%. The second classifier, Naive Bayes was also attempted and the model accuracy was 66% much lower than the accuracy model of J48. Because this reason rules generated by J48 were taken for analysis. Although J48 algorithm has resulted better accuracy, the attribute similarity classification; raises questions of misclassification. 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 requirement strategy and can provide a base and guidance for policy makers.

It is recommended for the blood bank service to plan strategically and collect blood from the group identified as less vulnerable by results of this study. This would maximize the collection of safe blood to be optimal.

References

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

[2] Bigus J. (1996). *Data Mining with Neural Networks: Solving* Business Problems- from Application Development to Decision Support. Mc Graw-Hill: New York.

[3] Butch S. H. (2002) Computerization in the transfusion service. Vox Sanguinis., 83 (suppl 1), 105-110.

[4] Dhingra N. (2016). Screening Donated Blood for Transfusion-Transmissible Infections: World Health Organization. Available at: http://www.who.int/bloodsafety/makingsafebloodavailableinafricastatement.pdf. Accessed August 2016.

[5] The Ethiopian Red Cross Society (2010). National Blood Bank Service Highlights Blood a Gift for Life.

[6] Shyamsundaram and Santhanam. T. (2010). Application of CART Algorithm in Blood Donors Classification PG and Research Department of Computer Science, DG Vaishnav College, Chennai-600106, Tamil Nadu, India.

[7] Belay T (2002). 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. Unpublishd Masters thesis Addiss Ababa University.

[8] Baye Gelaw and Yohans Mengistu (2002).. The prevalence of HBV, HCV and malaria parasites among blood donors in Amhara and Tigray regional states.

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

[10] Blood Safety Indicators (2009). World Health Organization. Geneva.

[11] Deogan (2011). Data Mining: research Trends, Challenges, and Applications [database on the Internet]. <http://citeseer.nj.nec.com/deogun97data.html> [Accessed on February, 21, 2016].

[12] Piatetsky-Shapiro G. (2000) Knowledge Discovery in Databases: 10 Years After. SIGKDD Explorations. Online. Retrieved from http://www.kdnuggets.com/gpspubs/sigkdd-explorations-kdd-10-years.html. Accessed March 15, 2016

[13] Han Ja K, Micheline (2001). Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.

[14] Last, Mark, Maimon, oded, and Kandel Abraham (2016). Knowledge Discovery in Mortality Records: Aninfo-fuzzy Approach. Retrieved from http://www. csee.usf.edu/ softec/med_ dm3.pdf. Accessed May 16, 2016.

[15] Fayyad U, Piatetsky-shapiro, G. and Smyth, Padharic (1996). From Data Mining to Knowledge Discovery in Databases.

[16] Helen T. (2003). Application of Data Mining Technology to Identify Significant Patterns in Census or Survey Data. Unpublished Masters Thesis Addis Ababa University, Addis Ababa.

[17] Tesfaye, Hintsay. (2002). Predictive Modeling Using Data Mining Techniques In Support to Insurance Risk Assessment.

[18] Cabena P. Discovering (19980. Data Mining - From Concept to Implementation, Prentice Hall, New Jersey.

[19] Thearling K. (2003). An introduction to Data Mining. Retrieved from http://www3.shore.net/~kht/text/dmwhite.pdf. Accessed March 18, 2016.

[20] Chapman P. (1999). CRISP-DM 1.0 Step-by-step Data mining Guide SPSS Inc., U.S.A CRISPWP-0800.

[21] Berry Mal, G. (1997). Data Mining Techniques: For Marketing, Sales and Customer Support. New York. John Wiley and Sons, Inc.

[22] Levin Na Z, Jacob, (1999). Data Mining. Available Retrieved from http:www.urbanscience.com/Data Mining.pdf

[23] Witten Ihaf, Eibe (2000). Practical Machine Learning Tools and Techniques with Java Implementations. USA: Academic Press.