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

CONSTRUCTING A PREDICTIVE MODEL FOR OCCURRENCE OF TUBERCULOSIS: THE CASE OF MENELIK II HOSPITAL AND ST. PETERS TB SPECIALIZED HOSPITAL

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
TEKETEL MULUGETA

MAY 2013

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

CONSTRUCTING A PREDICTIVE MODEL FOR OCCURRENCE OF TUBERCULOSIS: THE CASE OF MENELIK II HOSPITAL AND ST. PETERS TB SPECIALIZED HOSPITAL

BY
TEKETEL MULUGETA

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

Advisors:

MARTHA YIFRU (PhD)

MULUGETA BETRE (MD, MPH)

MAY 2013

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

_____

TEKETEL MULUGETA

MAY 2013

The thesis has been submitted for examination with my approval as university Advisors

_____

Martha Yifru (PhD)

MAY 2013

_____

Mulugeta Betre (MD, MPH)

MAY 2013

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

CONSTRUCTING A PREDICTIVE MODEL FOR OCCURRENCE OF TUBERCULOSIS: THE CASE OF MENELIK II HOSPITAL AND ST. PETERS TB SPECIALIZED HOSPITAL

BY
TEKETEL MULUGETA
MAY 2013

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

Name and Signature of Members of the Examining Board

Name		Signature	Date
_____	**Chairman**	_____	_____
_____	**Advisor**	_____	_____
_____	**Advisor**	_____	_____
_____	**Examiner**	_____	_____
_____	**Examiner**	_____	_____

ACKNOWLEDGMENTS

First of all, I would like to thank my almighty God for his unreserved provision during this research work.

I am deeply grateful to my advisor Martha Yifru (PhD) for her unreserved advice, keen insight, guidance and constructive comments I received throughout my thesis work. I also would like to express my heartfelt gratitude to my advisor Mulugeta Betre (MD, MPH) for his all rounded guidance and support in the course of writing this thesis.

I would like to thank managements and staff of Menelik II hospital and St. Peters TB specialized hospital for allowing me to access patients' medical history and assisting me whenever I needed their cooperation. I would like to thank especially Addisu Admasu (MD, MPH), Ato Abay Zewde, Ato Derege Asefa, Sister Memuna Yesuf, Sister Asnakech Tizazu, Sister Emebet Kuraze, W/ro Birhane Sisay, Ato Anbesaw Emyu, Ato Fiseha Abebe and W/ro Buzayehu Kassa, who are the staff member of the hospitals.

My special thanks goes to my best friends Ato Wondwosen Yakob (Nunu) for his support during data collection, data encoding and editing of this thesis, and Ato Fikre G/Hiwet, Ato Wondifraw Getachew and W/rt Rahel Moges for their support during data collection and data encoding.

My special thanks also goes to the health informatics coordinator W/rt Meseret Ayano for her unlimited support and helps whenever needed.

Finally, I would like to express my appreciation to my family and my best friends Ato Siyoum Taddese, W/ro Tigist Mulugeta, Ato Yosef Solomon, Ato Girma H/Gebrel, Ato Getachew Mekonnen, Ato Misganaw Taddese, Ato Dawit Ayele, Ato Fikerlenin Kassahun, Ato Moges Negede, Ato Moges Demlew and W/rt Abeba Belete for their love, understanding and unlimited support during this study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

ACF	Active Case Finding
AFB	Acid Fast Bacilli
AIDS	Acquired Immunodeficiency Syndrome
ARFF	Attribute Relation File Format
ANN	Artificial Neural Network
CFS	Correlation-based Feature Subset Selection
CRISP	Cross Industry Standard Process
CSV	Comma Separated Value
DOTS	Direct Observable Therapy
DM	Data Mining
EDA	Exploratory Data Analysis
EHR	Electronic Health Record
FMOH	Federal Ministry of Health
FN	False Negative
FP	False Positive
HIV	Human Immunodeficiency Virus
KBS	Knowledge Based System
KDD	Knowledge Discovery in Databases
LTBI	Latent Tuberculosis Infection
MDR-TB	Multi-Drug Resistant TB
MLP	Multilayer Perceptron
MMH	Maximum Marginal Hyperplane
M.TB	Mycobacterium Tuberculosis
PACF	Periodic Active Case Finding

PCR	Polymerase Chain Reaction
PTB	Pulmonary Tuberculosis
RIPPER	Repeated Incremental Pruning to Produce Error Reduction.
ROC	Receiver Operating Characteristics
SEMMA	Sample, Explore, Modify, Model and Assess
SVO	Support Vector Machine
SPSS	Statistical Package for Social Science
SMO	Sequential Minimal Optimization
TB	Tuberculosis
TN	True Negative
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSTRACT

Background: Tuberculosis is a disease of poverty affecting mostly young adults in their most productive years. In Ethiopia, TB is a disease of major public health problem. Early identification and isolation of TB cases is critical to prevent further transmission, morbidity and mortality caused by TB. Data mining has a potential to indentify hidden knowledge from huge datasets. It is possible to use data mining algorithms for analysis and predicting the TB status of patients.

Objective: The goal of this research was to apply data mining techniques for predicting the TB status of patients. Specifically, identify the determinant attributes of TB status of patients, build best prediction model and finally develop a prototype graphical user interface.

Methodology: A hybrid data mining process model that involved six steps is followed. This study considers a total of 10,031 records from Menelik II and St. Peters TB specialized hospitals patients' data and 15 attributes for predicting the TB status. Descriptive data analysis, visualization and statistical summary were implemented to gain understanding of the data. Handling of missing values and data transformation were done to prepare the dataset for experimentation. The mining algorithms used are decision tree, naïve bayes, support vector machine and artificial neural network. To evaluate the models performance 10-fold cross validation and confusion matrix are used.

Results: The result of the experiments with all and selected attributes showed that performance of J48, Sequential minimal optimization and Multilayer perceptron were better with all attributes than best selected attributes, whereas naïve bayes classifier performance increased with selected attributes than all attributes. The results of the experiments show the performance of mining algorithms decreases as the amount of training increases.

The best selected model to predict the TB status of patients in this study was generated by J48 decision tree with all attributes. The accuracy of this model is 95.24%. Graphical user interface prototype was designed using the ten rules from J48 decision tree.

Conclusion: The results achieved from this research indicate that data mining is useful in bringing relevant information from large and complex patients' dataset, and we can use this information for predicting TB status and decision making. The most important attributes that determine the TB status of the patients are shortness of breath, chest pain, cough, weight loss, loss of appetite, night sweats and HIV test results.

CHAPTER ONE
BACKGROUND

1.1. Introduction

Healthcare industry has experienced a production of innovations aimed at enhancing life expectancy, quality of life, diagnostic and treatment options, as well as the efficiency and cost effectiveness of the healthcare system(1). Healthcare data is generated exponentially on daily basis and this data is a key source to be analyzed and processed to extract hidden information and knowledge. It is very important that we could perform exploratory and predictive analysis on these data and put them in valuable and useful knowledge for health care staff.

Ethiopia experiences a heavy burden of diseases mainly attributed to communicable infectious diseases. Tuberculosis (TB) is one of these communicable diseases caused by a bacterium called Mycobacterium tuberculosis. The bacteria usually attack the lungs, but TB bacteria can attack any part of the body. If not treated properly, TB disease can be fatal. TB is spread through the air from one person to another when a person with active TB disease of the lungs or throat coughs, sneezes, speaks, or sings. People nearby may breathe in these bacteria and become infected. (2)

Tuberculosis is a disease of poverty affecting mostly young adults in their most productive years. The vast majority of TB deaths are in the developing world. 1.7 million people died from TB (600,000 women) in 2009, including 380,000 people with HIV, equal to 4,700 deaths a day. Reductions in the burden of disease achieved to date followed 15 years of intensive efforts to improve TB care and control. Between 1995 and 2009, a total of 41 million TB patients were successfully treated by the direct observable treatments (DOTS) program, and up to 6 million lives were saved including 2 million women and children (3). Ethiopia is now ranked seventh of 22 countries with the highest TB burden countries, which collectively contribute 80% of the global TB burden (4).

Data mining (DM) techniques are very popular for solving various problems. As a brief description, data mining is a mechanism for obtaining patterns from an existing dataset. Those extracted patterns are used to interpret the new or existing data into useful information (5). DM has a potential to indentify hidden knowledge from huge datasets. Many researchers apply data

mining to explore hidden pattern from electronic medical records. It is possible to use DM algorithms for analysis and predicting the prevalence of positive TB case. Using a model generated by data mining for early detection of positive TB case is economical and also improves the prevention of this disease by providing useful information for informed decision making.

As we have seen in the above paragraphs TB is a high burden disease in the developing countries and the early detection of this disease improves the prevention of TB cases. Therefore, the goal of this research was to apply the data mining techniques for predicting the TB status of patients from patients' dataset. This study used data mining technique for developing best model for predicting the TB status to improve the prevention of TB in our country.

The research outcome contributes to identify the important patterns of TB datasets, develop best prediction model which can be used for TB diagnosis, and build graphical user interface prototype for the discovered model. The result of this study used as a base for other researchers to develop knowledge based system (KBS) for prediction of TB status of patients.

1.2. Statement of the Problem

In Ethiopia, TB is a disease of major public health problem. Unfortunately, due to low health service coverage and poorly developed health information system, the actual magnitude of the TB problem has not been accurately determined (6). However, in 2007, the estimated incidence was 314 per 100,000 population; smear positive cases were 152 per 100,000 population; the prevalence of all forms of TB was 546 per 100,000 population; and mortality rate was 73 per 100,000 population (7). Ethiopia is ranked seventh of the 22 TB burden countries. According to the US global health policy 2011 report, the number of people living with TB was 200,000 (8). One of the main problems in TB prevention was the low detection rate. According to world health organization (WHO), there were 5.8 million notified cases of TB in 2009, the case detection rate of 63% (ranged 60–67%), up from 61% in 2008. The most important measure to prevent TB transmission is to maintain a high index of suspicion for TB. Early identification and isolation of TB cases is critical to prevent further TB transmission, morbidity and mortality (4). The existence of disease burden, low detection rate, and high death rate of TB at a national level make this research important.

On the other hand, the amount of data stored in medical databases has increased exponentially with time. The technological advancement resulted in the management of huge computerized data acquisition and storage of databases which contain hidden knowledge that can be used for decision making. It is impossible and time consuming to unravel this knowledge without incorporating data mining methods. Application of DM methods to discover hidden knowledge from medical record has been logical.

Different researches have been done on the prediction of TB abroad and in our country, such as Tamer U. (5) with title "Prediction the Existence of Mycobacterium Tuberculosis on Patients by Data mining Approach" has tried to apply different models on the data which are Bayesian Network, Multilayer perceptron, JRip, PART, RSES (a rough set algorithm implementation) and ANFIS (adoptive network implementation). The finding of the study indicate that ANFIS is an accurate and reliable methods comparing to Bayesian Network, Multilayer Perceptron, PART, JRip and RSES methods for classification of TB patients.(5)

Another research conducted by Asia T, et al. (9) with a title "Estimating the Statistical Significance of Classifiers used in Prediction of TB" has shown that support vector machine (SVM) is the best classifier. For their dataset there is no much significant difference in the accuracies of SVM and Decision trees. SVM proves to be with 99.07% accuracy whereas decision tree with 99% (9).

The third research done by Ali A.E, et al. (10) with a title "Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network" has shown that artificial neural networks can be used to predict active pulmonary TB using clinical and radiographic information. The finding of the study has indicated that General Regression Neural Network significantly performed with 94.76% +/- 2.8 (10). The above three researches indicates the applicability of classification mining algorithms in predicting the TB status and using many more mining algorithms help the researchers to select best mining algorithm for the given dataset.

The fourth research conducted in our country by Asia Nesredin (11) with a title "Mining Patients Data for Effective Tuberculosis Diagnosis: In case of Menelik II hospital" is used as a base to build the proposed study. In her study, clustering and classification data mining techniques were used to develop the model. For the clustering operation K-means algorithm was used and for

classification operation J-48 decision tree and naïve Bayes were used. Finally a model with 85.93% accuracy was developed using J48 Decision Tree (11). Even though, this research has generated a model for predicting TB cases, still there are certain gaps which are not addressed and need a thorough investigation. These gaps are clearly presented in the paragraphs below.

The first gap identified on the previous study was related to the use of very small dataset size to run the experiments. In data mining, the size of dataset used can have its own impacts on the accuracy and representativeness of the generated model.

The second problem on the previous study is, on the kind of mining algorithms used to develop the model. In that study only J48 decision tree and naïve bayes classification algorithm as well as K-mean clustering algorithm were used to develop the model. The best selected model has attained an accuracy of 85.93% entailing that there is a possibility of making 14.07% wrong decision. But if many more mining algorithms and large datasets were used, the resulting model might have a better accuracy than the previous study.

The third problem on the previous study is, selection of attributes used to develop the predict model. In that study irrelevant attribute treatment started year was used as one of best attribute and two important clinical symptoms such as loss of appetite and shortness of breath were not included in the previous study.

The last identified gap is on the usability of the research finding by the domain users to address their problems. Even though, the study has developed a predictive model, it does not address the existing problems of domain users.

Therefore, this research attempted to filled these gaps by taking the following major actions:

- The previous study has used only 7069 records from Menelik II hospital patient's dataset. So, the proposed study considered additional dataset from St. Peter's TB specialized hospital in addition to Menelik II hospital dataset. As a result more accurate and representative model was developed.
- This study exclude treatment started year attribute and includes two new attributes namely loss of appetite and shortness of breath based on the suggestion from domain experts.

- The study used Artificial Neural Network and Support Vector Machines algorithms in addition to J48 Decision Tree and Naïve Bayes algorithms (which were also used in the previous study) due to their usage on related problem domain.
- The study compared the performance of these four classification algorithms as the dataset size increases and selected the best performing algorithms.
- In this study a prototype graphical user interface was built in order to motivate the model usage by domain users.

Consequently, it has been attempted to answer the following major research questions:

- What are the most determinant attributes which can predict TB status of patients?
- Which mining technique produces the most accurate predictive model while adding the dataset size?
- Which mining algorithm has produced best TB prediction model?
- What kind of prototype can best communicate the discovered model with the domain users?

1.3. Objectives

1.3.1. General Objectives

The main objective of the study is applying data mining techniques to predict tuberculosis status (positive or negative) of patients from Menelik II hospital and St. Peter TB specialized hospital patients datasets.

1.3.2. Specific Objectives

The specific objectives are to:

- Identify the determinant attributes affecting TB status of a patient.
- Apply classification algorithms to train, test and build the classifier models using small and large dataset size.
- Compare the models based on their performance and select the best model.
- Develop a prototype graphical user interface capable of communicating the model found with the domain users to effectively utilize the model.

1.4. Scope and Limitations of the Study

The scope of this research is limited to constructing predicting models for the status of TB using Menelik II hospital and St. Peters TB specialized hospital patients datasets. Moreover, the study is limited to build a prototype graphical user interface for a TB prediction and use classification data mining technique. The classification algorithms selected by the researcher in this study to develop prediction models are Decision Tree, Naïve Bayes, Artificial Neural Network and Support Vector Machine.

Besides the promising findings observed in this study, the following major limitations are recorded:

- This study was conducted in two government hospitals namely, Menelik II and St. Peters TB specialized hospitals due two shortages of time and financial constraint.
- There is a need for the development of knowledge based system. However, this study limited to developed predictive model and prototype graphical user interface due to time constraints.
- This study is limited to developing predictive model using classification mining techniques. However, there are other mining techniques which might show some interesting patterns or relationships in the selected dataset. This was happened due to limitation of time encountered by the investigator.
- This study used data encoders to encode important variables from paper based medical records to computer. However, the correctness of the encoded data did not check to ensure data quality. This was due to limitation of time.

1.5. Significance of the Study

Tuberculosis is a disease for which (in most cases) an effective treatment is available; the role of the research on TB should be to provide tools which facilitate delivery of quality diagnosis and treatment. Even though, there were effective treatments for more than 30 years, there has been a failure to control TB in most of the developing countries (12). This study attempted to develop useful predictive model for tuberculosis status of patients and developed a prototype graphical user interface for effective utilization of the predictive model, and hence physician can make use of this model in their day-to-day diagnosis of tuberculosis. The diagnosis and treatment of TB have been highly uncertain because of low case detection rate, poor quality in diagnostic

procedures, and lack of knowledge by community health workers. Therefore, decisions support system in diagnosis of tuberculosis disease would be fundamental.

Physicians can make use of the results of this study in order to make optimal decisions about the presence or absence of TB disease. Moreover, it adds quality to decision making process for quality health care service provisions to the society at large.

In addition, health institutions and interested researchers use the result of the study as a good guideline towards the application of data mining on tuberculosis to create a model and to construct KBS for TB prediction that can help fighting tuberculosis diseases and its complication.

1.6. Organization of Thesis

This research report has seven chapters. The first chapter deals with introduction, statement of the problem, objective, scope of the study, significant of the study and thesis organization.

In the second chapter, review of literatures on data mining was presented. Definition and common tasks of data mining, data mining methodology, data mining algorithms and application of data mining in electronic medical record is discussed.

In the third second chapter, literature review about tuberculosis has been given which includes review of definition, risk factors, epidemiology and diagnosis of active tuberculosis. The chapter reviews current situation of tuberculosis in Ethiopia. Finally, review of related works particularly application of data mining in healthcare sector especially in tuberculosis prediction was discussed.

The fourth chapter discusses research methodology, data understanding and preprocessing tasks. In the research methodology section detail discussions about project area, research deign, evaluation credibility of the model is presented, ethical consideration and tools used in this study. Data source and collection, description of data, and descriptive statistical summary of selected attribute are discussed. In this chapter how the major data preprocessing tasks were applied to the current data are shown. Data cleaning and transformation was conducted for preparation of dataset. In the evaluation of credibility section details discussion presented about k-fold cross validation and confusion matrix.

The fifth chapter deals with experimentations and model selection. Experimental setup and attribute subset selection was presented. In this chapter, building of model with J48, naïve bayes, SMO and MLP, the validation of model with 10-fold cross validation, performance evaluation of algorithms with attribute subset selection and increasing dataset size, and interpretation of the result of the experiments are the major concerns. Finally, a comparison of models, selection of the best model, generated rules from best model and development of a prototype graphical user interface was done.

The sixth chapter covers conclusions and recommendations.

CHAPTER TWO

DATA MINING CONCEPTS

2.1. What is Data Mining?

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable one. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides (13, 14). Data mining refers to extracting or mining knowledge from large amounts of data (15). Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviours and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that were traditionally too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations (14).

Technological developments that aid to collect and store vast quantities of data have enabled organizations to capture and cumulate huge amount of data in their databases, within which, large amount of valuable information is buried (16). It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% (17). As the volume of data increases, the proportion of information in which people could understand decreases substantially. This reveals that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which results with increasing information gap. Consequently, people begin to realize this bottleneck and to look into possible remedies. With the aim to fix the gap between the generation of data and rate of understanding of people about it, data mining can serve as a principal tool.

Data mining has been defined in almost as many ways as there are authors who have written about it. According to Berry and Linoff (18), data mining is the process of exploration and analysis by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. Data mining usually makes sense when there is large amount of data. For 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 responsible for different

tasks of data mining such as classification, clustering, prediction, and association. The need for bulky data can be explained by a couple of reasons. Primarily, in the case of small databases, it is feasible to capture appealing trends and relationships by introducing traditional tools such as spreadsheets and database query. The second reason is that most data mining tools and algorithms demand large amount of training data (data used for building a model) in order to generate unbiased models. The rationale is simple and straightforward, small training data results in unreliable generalizations based on chance patterns.

According to Witten and Frank (17) data mining is valuable to discover implicit, potentially useful information from huge data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns. The opportunity for the application of data mining has increased significantly as databases grew extremely and new machine with searching capabilities evolved.

2.2. Common Tasks of Data Mining

It is convenient to categorize data mining into types of *tasks*, corresponding to different objectives for the person who is analyzing the data. The book Principles of Data Mining categorizes data mining tasks as follows (19):

> ***Exploratory Data Analysis (EDA)*** *As the name suggests, the goal here is simply to explore the data without any clear ideas of what we are looking for. Typically, EDA techniques are interactive and visual, and there are many effective graphical display methods for relatively small, low-dimensional data sets.*
>
> ***Descriptive Modeling*** *The goal of a descriptive model is to describe all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (density estimation), partitioning of the p-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling). In segmentation analysis, for example, the aim is to group together similar records, as in market segmentation of commercial databases. Here the goal is to split the records into homogeneous groups so that similar people (if the records refer to people) are put into the same group. This enables advertisers and marketers to efficiently direct their*

promotions to those most likely to respond. The number of groups here is chosen by the researcher; there is no "right" number.

Predictive Modeling: Classification and Regression: *The aim here is to build a model that will permit the value of one variable to be predicted from the known values of other variables. In classification, the variable being predicted is categorical, while in regression the variable is quantitative. The term "prediction" is used here in a general sense, and no notion of a time continuum is implied. So, for example, while we might want to predict the value of the stock market at some future date, or which horse will win a race, we might also want to determine the diagnosis of a patient, or the degree of brittleness of a weld.*

Discovering Patterns and Rules: *The three types of tasks listed above are concerned with model building. Other data mining applications are concerned with pattern detection. One example is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest. Another use is in astronomy, where detection of unusual stars or galaxies may lead to the discovery of previously unknown phenomena. Yet another is the task of finding combinations of items that occur frequently in transaction databases (e.g., grocery products that are often purchased together). This problem has been the focus of much attention in data mining and has been addressed using algorithmic techniques based on association rules.*

Retrieval by Content: *Here the user has a pattern of interest and wishes to find similar patterns in the data set. This task is most commonly used for text and image data sets. For text, the pattern may be a set of keywords, and the user may wish to find relevant documents within a large set of possibly relevant documents (e.g., Web pages). For images, the user may have a sample image, a sketch of an image, or a description of an image, and wish to find similar images from a large set of images. In both cases the definition of similarity is critical, but so are the details of the search strategy.*

The following are the three most common data mining tasks:

2.2.1. Classification and prediction

Classification is a process of finding a set of models (or functions) that uncover and distinguish 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). The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formula, and neural network. Classification can be used for predicting the class label of objects. (15)

2.2.2. Association Rule Discovery

One of the tasks of DM is association rule mining. Association rule mining finds interesting association or correlation relationships among a large dataset. With a massive amounts of data continuously being collected and stored, many industries became interested in mining association rules from their datasets. The discovery of interesting association among huge amount of business transaction records can help in many business decision making processes. Association rules are in the form of "If antecedent, then consequent," together with a measure of the support and confidence associated with the rule. (15)

2.2.3. Clustering

Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized. (20)

2.3. Data Mining Methodology

Data mining and knowledge discovery projects involve a complex process of identifying, understanding, and transforming data modeling and evaluation efforts. Naturally, any models that are useful would need to be deployed before making an impact in practice. There are several data mining process frameworks that seek to organize these activities (16). Many data mining process methodologies are available. However, the various steps do not differ much from methodology to methodology. The most common data mining methodology are the following:

2.3.1. CRISP-DM

A widely cited industry framework is the Cross-Industry Standard Process for Data Mining (CRISP-DM) model. CRISP-DM offers a general model for data/text mining projects, highlighting the key tasks involved. According to the CRISP-DM framework, the life cycle of a knowledge discovery project consists of six phases, but the sequence of the phases is not strictly applied. Moving back and forth between different phases is always required. The process is iterative because the choice of subsequent phases often depends on the outcome of preceding phases. The life cycle begins with business understanding to ground the overall aims of the project, and then moves to data understanding to identify potential inputs and outputs, data quality issues, and potential privacy or security concerns. The third phase, data preparation, involves the extraction of relevant data for a particular modeling effort, data quality assurance, and any transformations required for specific modeling techniques. Typically, the data preparation tasks account for the majority of effort in a data mining project. (13)

The fourth phase, data modeling, is the central focus of any knowledge discovery effort and consists of the construction of models based on a variety of techniques, with evaluations (the fifth phase) conducted for all modeling techniques. The final step is deployment so that useful models can be embedded in information systems to support decision-making activities. (14)
The next figure show the CRISP-DM methodology knowledge discovers steps:

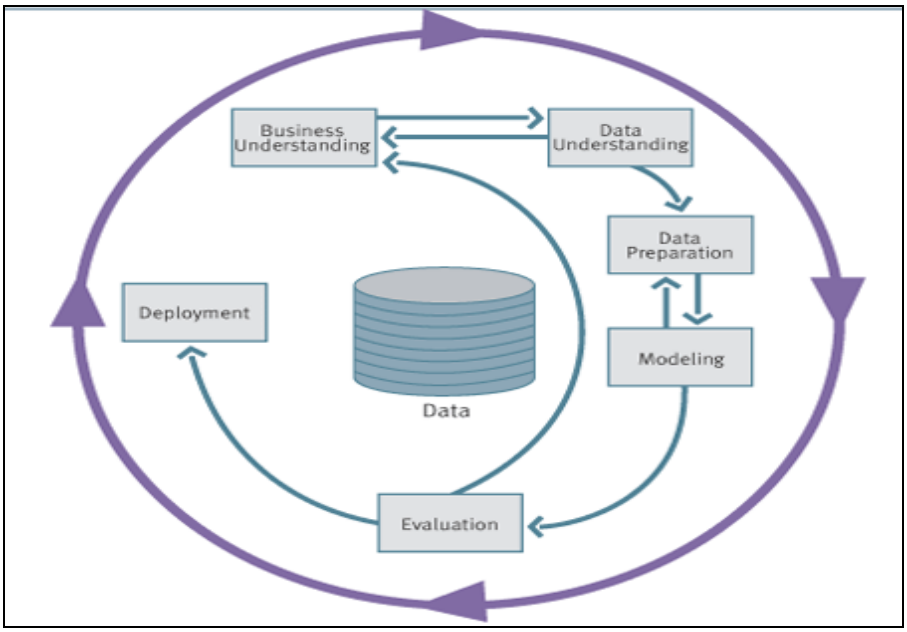


Figure 2.1 CRISP-DM process model

2.3.2. KDD Process

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from large and complex data sets (16). KDD is an automatic, exploratory analysis and modeling of large data repositories (21). The KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database (16). There are considered five stages, presented in figure 2.2:

1. Selection – This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
2. Pre processing – This stage consists on the target data cleaning and pre processing in order to obtain consistent data.
3. Transformation – This stage consists on the transformation of the data using dimensionality reduction or transformation methods.
4. Data Mining – This stage consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction)
5. Interpretation/Evaluation – This stage consists on the interpretation and evaluation of the mined patterns.

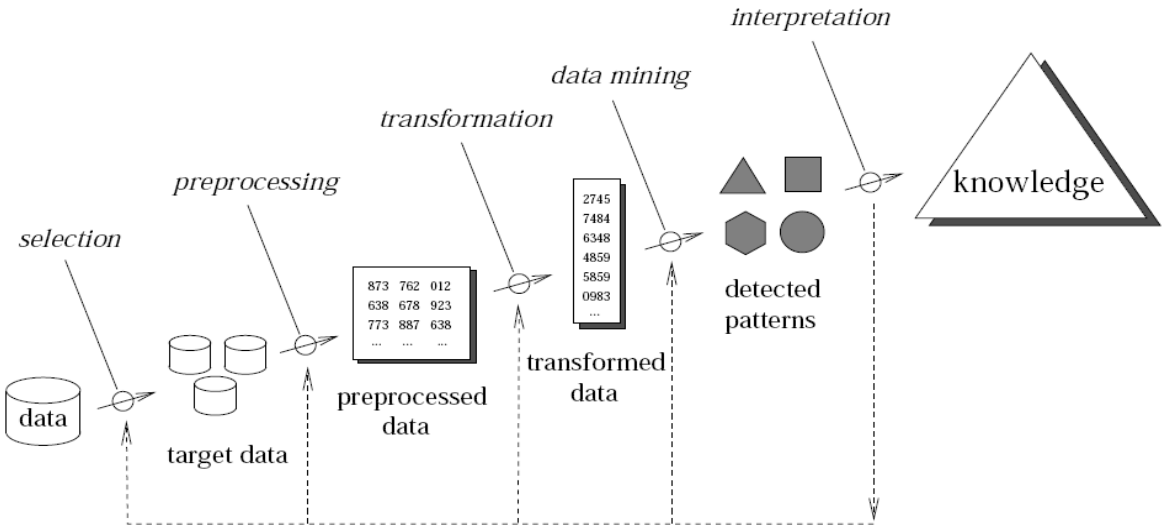


Figure 2.2 KDD process model

2.3.3. SEMMA

Statistical Analysis Software (SAS) developed a data mining analysis cycle known by the acronym SEMMA. This acronym stands for sample, explore, modify, model, assess. Beginning with a statistically representative sample of your data, SEMMA intends to make it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and finally confirm a model's accuracy. A graphic representation of SEMMA is given in Figure 2.3. By assessing the outcome of each stage in the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. (22)

The followings are steps in SEMMA process:

1. Sample: the first step in is to create one or more data tables by sampling data from the data warehouse. Mining a representative sample instead of the entire volume drastically reduces the processing time required to obtain business information.
2. Explore: after sampling the data, the next step is to explore the data visually or numerically for trends or groupings. Exploration helps to refine the discovery process. Techniques such as factor analysis, correlation analysis and clustering are often used in the discovery process.

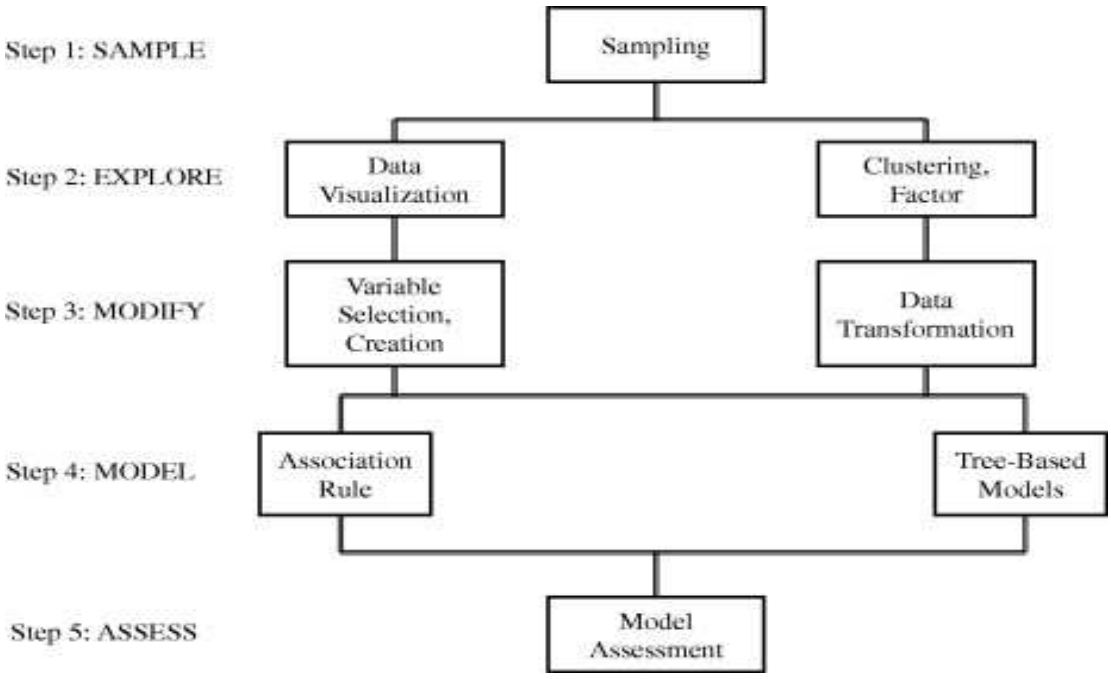


Figure 2.3: SEMMA process model

3. Modify: modifying the data refers to creating, selecting, and transforming one or more variables to focus the model selection process in a particular direction, or to modify the data for clarity or consistence.
4. Model: creating a data model involves using the data mining software to search automatically for a combination of data that predicts the desired outcome reliably.
5. Assess: the last step is to assess the model to determine how well it performs. A common means of assessing a model is to set aside a portion of the data during the sampling stage. If the model is valid, it should work for both the reserved sample and for the sample that was used to develop the model. (22)

The SEMMA approach is completely compatible with the CRISP approach. Both aid the knowledge discovery process. Once models are obtained and tested, they can then be deployed to gain value with respect to business or research application.

2.3.4. Hybrid Data mining

The development of both academic particularly the KDD and industrial oriented (CRISP-DM and other) data mining models has led to the growth of hybrid models, i.e., models that combine the features and job of both. Hybrid model was developed mainly based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include introducing several new explicit feedback mechanisms and in last steps the knowledge discovered for a particular domain may be applied in other domains. The Hybrid DM (see figure 2.4) consists of six-step Knowledge Discovery Process (23).

According to Cios et al. (23) the description of the six steps follows:

1. Understanding of the problem domain: - This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology.
2. Understanding of the data*:* - This step includes collecting sample data and deciding which data, including format and size, will be 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.

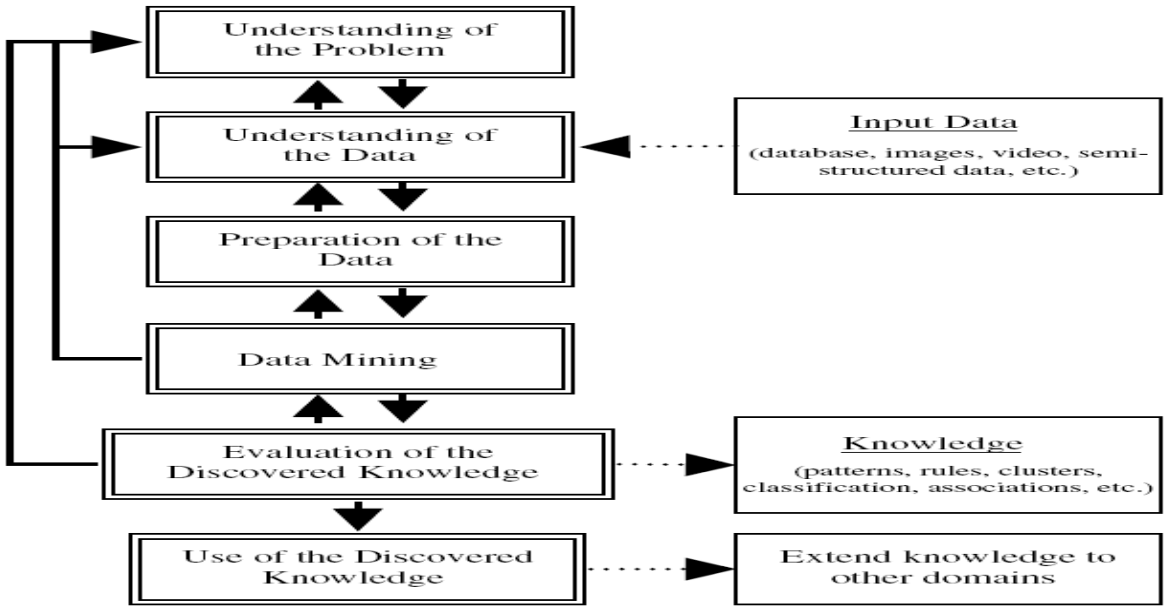


Figure 2.4 Hybrid Process Model

3. Preparation of the data: - This step concerns deciding which data will be 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 end results are data that meet the specific input requirements for the DM tools selected in Step 1.
4. Data mining: - Here the data miner uses various DM methods to derive knowledge from preprocessed data.
5. 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.
6. Use of the discovered knowledge: - This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.4. Mining Algorithms

According to Berry and Linoff (19), learning and understanding of different data mining techniques is essential for the following reasons:

- To take the advantage of a specific technique,
- To determine the best applicable technique for the problem at hand.
- To know the advantages and disadvantages of a technique.

It is apparent that no technique is applicable to all data mining problems. To determine the best technique suitable to the specific data mining problem, familiarity with the available techniques is necessary. There are many data mining classification algorithms available and commonly used in various applications. The researcher used decision tree, naïve bayes, artificial neural network and support vector machine mining algorithms in this research.

2.4.1. Decision Trees

A Decision tree is a classifier expressed as a recursive partition of the instance space. A decision tree is "a flow-chart-like tree structure." (24). Decision tree divide the records in the database into subsets based on the values of one or more fields. This process will be repeated for each subset recursively until all the instances at each node fall in to a single class. The outcome of decision tree is a tree shaped structure that depicts a series of decision made at each step. Then these decisions are considered as rules for the task of classification (18).

Decision tree are powerful and widely used data mining method for classification and prediction (18). The strength and popularity of decision tree is due to the fact that in contrast to neural networks, it expresses the 'if-then' rules explicitly. In applications where the accuracy of a classification or prediction of unknown instances is the only thing that matters, that is, in cases where how or why the model works is not important, both neural networks and decision tree can perform a good job (17).

Decision tree methods are equally well suited at handling continuous and categorical variables. Categorical variables (take values in a pre-specified, finite set of possibilities), which pose problems for neural networks and statistical techniques, come ready-made with their own splitting criteria: one branch for each category. On the other hand, continuous variables (that take

numbers either real or integer values) are equally easy to split by picking a number somewhere in their range of values.

Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value. The resulting rule set can then be simplified to improve its comprehensibility to a human user, and possibly its accuracy (25).

Decisions tree algorithm have three parameters. These are: data partition (a complete set of training tuples and their associated class labels), attribute list (a list of attributes describing the tuples), and attribute selection method which select best discriminant attribute given tuples according to class. This method employs an attribute selection measure, such as information gain or gini index (15). The typical difference between information gain and gini index is that of gini index enforces the resulting tree to be binary where as information gain allows a multiway splits or it allows two or more branches to be grown from a node.

2.4.1.1. J48 Decision Tree Induction

The basic algorithms for decision tree induction is a greedy algorithm which constructs decision trees in a top down approach dividing each node recursively until a leaf node is encountered. The following algorithm shows how decision tree algorithms generate a decision tree from the given training data (15).

> ***Input:*** *The training samples, samples, represented by discrete-value attributes; the set of candidate attributes, attribute-list.*
> ***Output****: A decision tree.*
> ***Method****:*
> *1) Create a node N;*
> *2) If samples in D are all the same class, C then*
> *3) Return N as a leaf labeled with the class C;*
> *4) If attribute list is empty then*
> *5) Return N as a leaf node labeled with the most common class in samples;*

6) *Select test-attribute, the attribute among attribute-list with the highest information gain;*
7) *Label node N with test-attribute;*
8) *For each known value ai of test-attribute // partition the samples*
9) *Grow a branch from node N for the condition test-sample = ai;*
10) *let si be the set of samples in samples for which test-attribute = ai;*
11) *if si is empty then*
12) *Attached a leaf labeled with the most common class in samples;*
13) *else attach the node retured by Generate decision tree (si, attribute-list-test-attribute);*

2.4.2. Naive Bayesian

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered "naive" (15).

2.4.2.1. Bayes' Theorem

Let X be a data sample whose class label is unknown. In Bayesian terms, X is considered "evidence". Let H be some hypothesis, such that the data belongs to a specific class C. For classification problems, our goal is to determine P (H|X), the probability that the hypothesis H holds given the observed data sample X). In other words, we are looking for the probability that sample X belongs to class C, given that we know the attribute description of X. (15)

P (H|X) is the a posteriori probability of X conditioned on H. P(X) is the prior probability of X. Bayes theorem is useful in that it provides a way of calculating the posterior probability, P(H/X), from P(H), P(X), and P(X/H) (15). Bayes theorem is

$$P(H/X) = \frac{P(X/H)\,P(H)}{P(X)}$$

2.4.2.2. Naive Bayesian Classifier

The naïve Bayesian classifier works as follows (15):

1. *Each data sample is represented by an n-dimensional feature vector, X = (x1, x2, …, xn), depicting n measurements made on the sample from n attributes, respectively, A1, A2, … , An.*
2. *Suppose that there are m classes, C1, C2, … , Cm. Given an unknown data sample, X (i.e., having no class lable), the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is, the naïve Bayesian classifier assigns an unknown sample X to the class Ci if and only if*

   $P(Ci/X) > P(Cj/X)$ *for* $1 \leq j \leq m,\ j \neq i.$

   Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

   $$P(Ci|X) = \frac{P(X/Ci)\,P(Ci)}{P(X)}$$

3. *As P(X) is constant for all classes, only P(X|Ci) P(Ci) need to be maximized. If the class a priori probabilities, P(Ci) are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = . . . = P(Cm), and we would therefore maximize P(X|Ci). Otherwise we maximize P(X|Ci) P(Ci).*

4. *Given data sets with many attributes, it would be computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means that*

   $$P(X|Ci) = \prod_{K=1}^{n} P(xk/Ci).$$

   The probabilities P (x1|Ci), P(x2|Ci), … , P(xn/Ci) can easily be estimated from the training samples, where

A. *If Ak is categorical, then P(xk|Ci)= is the number of samples of class Ci in T having the value xk for attribute Ak, divided by freq (Ci, T), the number of sample of class Ci in T.*

B. *If Ak is continuous-valued, then the attribute is typically assume that the values have a Gaussian distribution with a mean μ and standard deviation _ defined by*

P(xk/Ci) = g(xk, μCi, σCi) = 1p2__ exp−(x − μ)2 2_2 ,

where g(xk, μCi , σCi) is the gaussian (normal) density function for attribute Ak, while μCi and σCi are the mean and standard deviation, respectively, given the values for attribute Ak for training samples of class Ci.

5. *In order to classify an unknown sample X, P(X|Ci) P(Ci) is evaluated for each class Ci. Sample X is then assigned to the class Ci if and only if*

P(X/Ci)P(Ci) > P(X/Ci)P(Ci) for 1 ≤ j ≤ m, j ≠ i.

In other words, it is assigned to the class Ci for which P(X|Ci) P(Ci) is the maximum.

2.4.3. Artificial Neural Network

The study of neural networks commonly referred to as neural networks has been motivated from recognition of the fact that biological learning processes are entirely different from the way the traditional digital computers process information (24). The biological brain is made up of very complex and massive webs of inter-networked neurons. Although the brain is “very complex, nonlinear and parallel information processing,” it is powerful enough to manage and control neurons in order to successfully undertake very complex classification such as pattern recognition, image recognition, perception and motor control several times faster than the latest digital computer, which is prevalent today (17, 25).

Although the human brain contains inter-networked networks of about 10^{11} neurons and each neuron connected approximately to 10^4 neurons, is quick enough to make complex decisions (24). According to Haykin (25), intellectuals think that the information processing abilities of biological neural systems must emanate from highly parallel processes operating on representations that are dispersed across a number of neurons. Therefore, one motivation for the

study of neural network is to simulate highly parallel information processing based on distributed representations.

Neural networks are considered to be black boxes because of their non-linear behaviors and are usually more complicated than other techniques (18). They added that training a neural network is a further challenge requiring setting with numerous parameters and the output of a neural network is not as easily understood by the user as the output seen by a decision tree tool. In spite of this, neural networks are proving their worth every day in a wide variety of practical application. Moreover, several algorithms have recently been developed for the extraction of rules from trained neural networks (15).

2.4.3.1. Structure of Neural Networks

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. 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. (15, 18, 25)

2.4.3.2. Classification of Neural Networks

Depending on the pattern of connectivity, two types of networks can be distinguished: feed forward and recurrent networks (25):

- ***Feed forward networks:*** are employed in situations when we can bring all of the information to bear on a problem at once, and we can present it to the neural network. In this type of network, the data flows through the network in one direction, and the answer is based solely on the current set of inputs. In other words, they have no feedback connection i.e. they have no connections through weights extending from the outputs of a layer to the inputs of the same or previous layer. Feed forward networks have no memory; their output is solely determined by the current inputs and the values of the weight.

- ***Recurrent networks:*** are networks with feedback connections. In some configurations, recurrent networks re-circulate previous outputs back to inputs; hence, their output is determined both by their current input and their previous outputs.

Neural networks could also be discussed in terms of their learning mechanisms. The learning rule is the very heart of a neural network. It determines how the weights are adjusted as the neural network gains experience. Many networks use some variation of the Delta rule (to search the hypothesis space of possible weight vectors to find the weights that best fit the training examples) for training. One type of the rule most widely used is back propagation.

Training algorithms can be classified as supervised and unsupervised training (24):

- ***Supervised training:*** is the process of learning in which the network learns from examples, i.e. a pair of input vector with target vector containing the desired output. The network is provided with instances whose class are known and let learn from these instances. The learning algorithm takes the difference between the correct output and the prediction of the neural networks so that the prediction next time would be closer to the correct answer. The neural network has to be given examples many times in order to learn and make correct prediction. The back propagation algorithm employs the supervised learning paradigm.
- ***Unsupervised training:*** is the process of learning in which the training set consists solely of input vectors. In this approach of training there are no target outputs and it is impossible to determine what the result of the learning process will look like. The network is simply provided with a number of inputs and the network organizes itself in such a way as to come up with its own classification of inputs.

The back propagation algorithm is the algorithm considered in this study since it is the most widely used algorithm in the study of neural networks. Often, back propagation algorithm comprises the following three steps (24).

- The input pattern is presented to the network whereby the input pattern is propagated through the network until they reach the output units. This forward pass would produce the actual or predicted output.
- Then the desired output would be given as part of the training set, so that the actual output can be subtracted from the desired output in order to give the error signal.

- In the last step the errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output. The connection weights are then adjusted and the neural network is said to have learned from an experience.

The above steps are repeatedly carried out for all examples in the data until the termination condition of the iteration is satisfied.

2.4.3.3. Multilayer Perceptron

A Multilayer Perceptron is the most known and most frequently used type of neural network. The signals are transmitted within the network in one direction: from input to output. That is, there is no loop; the output of each neuron does not affect the neuron itself. This architecture is called feed forward neural network. In the MLP structure, the neurons are grouped into layers. The first and last layers are called input and output layers respectively because they represent inputs and outputs of the overall network. The remaining layers are known as hidden layers (26).

Typically, a multilayer Perceptron consists of a set of sensory units or source nodes that constitute the input layer, one or more hidden layers of computational nodes and an output layer of computational nodes. The input signal propagates through the network in a forward direction on a layer-by-layer basis. Multilayer Perceptrons have been applied successfully to solve difficult and diverse problems by training them in a supervized manner (26).

2.4.4. Support Vector Machine

Support Vector Machines (SVMs) are a set of related methods for supervised learning, applicable to both classification and regression problems. Since the introduction of the SVM classifier a decade ago, SVM gained popularity due to its solid theoretical foundation. The development of efficient implementations led to numerous applications (23). SVM is a promising new method for the classification of both linear and nonlinear data; it uses a nonlinear mapping to transform the original training data into a higher dimension (15).

To explain SVM; if it is a two-class problem where the classes are linearly separable, an algorithm is implemented to find a special kind of linear models. Let xi $\in$ Rn, (i= 1, 2, 3… m) represents the vectors and yi $\in$ {1,-1}. The term f(*xi*) can be represented by a linear functions of the form by yi= *f*(*xi*):

$$f(xi) = (W.\ X) + b$$

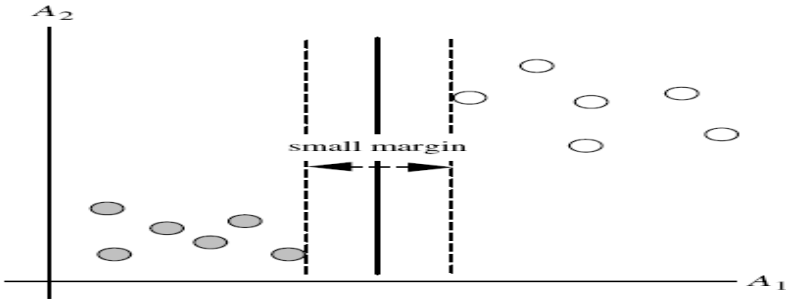


Figure 2.5: The 2-D training data are linearly separable.

Where W is a weight vector namely, W= {W1, W2, W3 . . . Wn} and b is a scalar, often referred to as bias (15). There is infinite number of hyperplane/ separating lines that could be drawn for classifying the two-classes (15). To find the optimal linear model or Hyperplane (n dimensions) that will have the minimum classification error on previously unseen tuples, SVM search for maximum marginal hyperplane (17).

Maximum marginal hyperplane is the one that gives the greatest separation between the classes (17). Figure 4.1 depict that hyperplanes that can correctly classify all of the given data samples. But larger margin are likely to be more accurate at classifying future data tuples than the hyperplane with the smaller margin.

During learning phases, SVM searches for the hyperplane with largest margin, which is the maximum marginal hyperplane (MMH) (15). When dealing with the MMH, this distance is the shortest distance from the MMH to the closest training tuple of either class (15). A hyperplane separating the two classes' decision boundary may be written as:

$$X=b+\Sigma\alpha_i y_i \mathbf{a(i).a}$$

Here, yi is the class value of training instance a(i); while b and αi are numeric parameters that have to be determined by the learning algorithms. a(i) and a represent the vectors. The vector a represent the test instances and a(i) are the training instances.

2.4.4.1. Sequential Minimal Optimization

Sequential minimal optimization (SMO) implements the SMO algorithm for training a support vector classifier, using polynomial or Gaussian kernels. Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default-note

that the coefficients in the output are based on the normalized data. Normalization can be turned off, or the input can be standardized to zero mean and unit variance. Pairwise classification is used for multiclass problems. Logistic regression models can be fitted to the support vector machine output to obtain probability estimates. In the multiclass case the predicted probabilities will be coupled pair wise. When working with sparse instances, turn normalization off for faster operation. SMOreg implements the sequential minimal optimization algorithm for regression problems (17).

2.5. Data Mining in Electronic Health Records

The Healthcare Information and Management Systems Society defines an Electronic medical records (EHR) system as a clinical information system, owned and operated by a healthcare delivery organization, that serves as the legal record of a patient encounter. (27)

Data mining can be described as the inductive approach to uncovering patterns in data. In the particular case of data mining on EHR, the process requires the application of machine learning techniques, or more precisely, knowledge discovery in databases. The use of data mining techniques could allow the classification of patients by risk category, or could help predict the length of hospitalization for a specific patient after a surgical procedure. The knowledge discovery process is as a multistep life cycle that begins with problem and data understanding, explicitly highlights the large effort necessary for data preparation, and then proceeds to modeling and evaluation. The final phase is the deployment of predictive models into existing systems. (13)

Machine learning uses an inductive approach to uncovering patterns in EMR's in order to search for mapping functions that relate a set of input data elements to corresponding output variables. Machine learning algorithms provide inductive, data-driven approaches to a wide variety of tasks, such as classification and prediction. Classification tasks try to group individual data entries into a known set of categories. Predictive tasks look into the future, applying a model to new data and making a qualitative classification or even a quantitative estimation. (13)

CHAPTER THREE

TUBERCULOSIS

3.1. What Is Tuberculosis?

In 1882, the German scientist Robert Koch identified a species of bacteria, *Mycobacterium tuberculosis (*M.TB or tubercle bacilli) as the cause of tuberculosis (8). Tuberculosis (TB) is bacterial disease caused by mycobacterium tuberculosis (and occasionally by mycobacterium bovis and mycobacterium africanum). These organisms are also known as tubercle bacilli or acid fast bacilli (AFB). Most commonly TB transmitted by air born spread of infectious droplets (12). The bacteria usually infects the lungs, but can involve any other organ (28).

There are two general stages of the disease: tuberculosis infection (or latent TB) and active tuberculosis. Individuals with tuberculosis infection are asymptomatic and not contagious, whereas individuals with active TB can be symptomatic and contagious (8). Usually the bacteria enter a dormant state (latent tuberculosis infection, LTBI), in which it remains for a period lasting from weeks to many years (28). Tuberculoses infection is necessary to develop TB, but overall only 10 percent of those with the infection ever develop active TB. Risk is higher for children and for people with human immunodeficiency virus (HIV) and other disorders that impair immunity (8).

The principal source of infection is the patient with TB of the lung, or pulmonary TB (PTB) who is coughing. This person is usually sputum smear positive. Coughing produces tiny infections droplet nuclei. Droplet nuclei can be spread in to the air by talking, sneezing, spitting and singing, and can remain suspended in the air for long periods. (29)

Pulmonary TB is the most common and leads to the destruction of lung tissue and frequently death if untreated. Symptoms include weakness, fever, chest pain, cough, and when a small blood vessel is eroded, bloody sputum. Active TB can also occur in other parts of the body, with the brain (TB meningitis) being the most serious. TB outside the lungs is more likely to occur among children and people with HIV. (8)

3.2. Tuberculosis Infection

Tuberculosis infection occurs when a person carries the tubercle bacilli inside the body, but the bacteria are in small numbers and are dormant. These dormant bacteria kept under control by the body's defense system and do not cause disease. This stage is known as latent TB. Many people have TB infection and are well. When our body defense system get weak because of various causes, the tubercle bacilli in the body will start to multiply and become numerous enough to overcome the body's defense and cause active tuberculosis. (28, 29)

3.3. Risk of TB Infection

An individual risk of infection depends on the extent of exposure to droplet nuclei and his/her susceptibility of infection. The risk of infection of a susceptible individual is high with close, prolonged, indoor exposure to a person with sputum smear positive PTB. The risk of transmission of infection from a person with sputum smear negative PTB is low, and even lower from someone with extra-pulmonary TB. Infection with mycobacterium tuberculosis can occurs at any age. Once infected with TB, a person can stay infected for many years, probably for life. Infected persons can develop TB at any time. (29)

3.4. Epidemiology of TB

Tuberculosis has long been recognized as major cause of morbidity and mortality (1). Tuberculosis infects nine million people and kills almost two million people each year, primarily in resource limited settings (30). TB has infected one third of the world population. In 2000 there were an estimated 8.5 million new cases of TB worldwide, 95% of TB cases and 98% of TB deaths are in the developing countries. 75% of TB cases in developing countries are in the economically productive age group (15-50 years). In 2000, sub-Saharan African had the high TB incidence of cases (6%) (29).

In Ethiopia, TB is a disease of major public health problem. Unfortunately due to low health service coverage and poorly developed health information system, the actual magnitude of the TB problem has not been accurately determined (6). The National Tuberculosis and Leprosy Control Program *(*NTLCP*)* estimates that the annual number of new cases amount to about 90,000 of which about 45% are open pulmonary tuberculosis cases. Some of the main reasons suggested for the widespread of pulmonary tuberculosis are HIV infection, neglect of

tuberculosis program, rapidly growing slums with crowded living conditions, lack of access to modern health care and deficient medical services (7). According to the Global Report 2009 by WHO, Ethiopia ranked as 7th among the TB high burden countries in the world, with an estimated incidence of all forms of TB of 378 new cases/100,000 population/year and 163 new smear positive cases/100,000 population/year; the estimated prevalence of all forms of TB is 579/100,000 population and 286/100,000 population as far as smear positive TB is concerned (31). According to Federal Ministry of Health (FMOH) manual, in 2007 the estimated incidence was 314 per 100,000 population, smear positive cases was 152 per 100,000 population, the prevalence of all form of TB was 546 per 100,000 population, and mortality rate was 73 per 100,000 population (7).

3.5. Diagnosis of Active TB

Current practice for diagnosis of active TB includes review medical history, clinical symptoms, laboratory tests, chest x-ray, and tuberculin skin test findings. Although these technologies have been generally adequate, their deficiencies have grown in the face of rising drug resistance and the need to prevent the rapid spread of TB among patients who have HIV or who live in congregate settings. (8)

The initial diagnostic laboratory test is the sputum smear in which a sample of the patient sputum is stained with a dye for AFB and is examined under a light microscope. However, only 50 to 80 percent of patients with active TB have positive sputum smears, and the rate may be even lower for people with HIV. Hence, negative smears cannot be used to rule out TB. Chest x-rays are also used to detect signs of the presence of TB in the lungs or the damage caused by the disease (8). In the case where diagnosis doubt persists, sputum culture may be useful if suitable facilities are available. In populations with a high prevalence, the tuberculin skin test also used to identify patients infected with M.TB. (11, 29)

Newer diagnostic technologies under study include a genetic technique called polymerase chain reaction (PCR) to amplify the amount of deoxyribonucleic acid (DNA) material specific to tubercle bacilli. PCR-based diagnosis has repeatedly produced results in research conditions within 48 hours. (8)

3.6. Tuberculosis Control Program

WHO has declared that TB is global emergency, because TB is out of control in many parts of the world. TB control program in the world has developed between 1948 and 1963. The new realistic TB control program had been launched in 1964. However TB did not decrease as expected. The goals of TB control program are to reduced mortality, morbidity, and disease transmission until TB no longer poses a threat to public health. The targets of the program are to cure 85% of the sputum smear positive case detected and to detect 70 % of existing cases of sputum smear positive PTB. (29)

The successful TB control depends on health care workers treating TB patient with national TB program framework and full implementation of the DOTS strategies. This means ensuring the accurate diagnosis and effective treatment of all TB patients. DOTS is way to ensure patients adherence to treatment. This means that someone supports the patient during the course of treatment and watches the patient swallow the tablets. (29)

DOTS strategy expanded and strengthen TB control program. DOTS increase the global new smear positive cases from 11% in 1995 to 45% in 2003 and could reached 60% by 2005. The WHO has developed new strategy, which builds on the success of DOTS and addresses unmet needs, the global plan to stop TB (2006-2015). The goal of stop TB strategy was to dramatically reduce the global burden of TB by 2015 in line with the millennium development goals and the stop TB partnership targets. (30, 32)

3.7. TB Situation in Ethiopia

Ethiopia is located in the North Eastern part of Africa, also known as the Horn of Africa. The total population of Ethiopia is 85 million, according to the national census conducted in 2007 by Central Statistical Authority. Administratively the country is divided in 9 Regional States and two City Administrative Councils (Addis Ababa and Dire Dawa).

Tuberculosis have been recognized as major public health problems in Ethiopia more than half a century ago. The effort to control tuberculosis began in the early 1960s with the establishment of TB centers and sanatorium in three major urban areas in the country. In 1976, in order to address effectively the TB challenge, the Central Office of the National Tuberculosis Control Programme (NTCP) was established. The challenge to control TB in Ethiopia is the emergence of multi-drug

resistant TB (MDR-TB). The data from drug sensitivity testing survey conducted in the country between 2003 and 2006 shows that levels of MDR-TB are: 1.6% and 11.8% in new cases and retreatment cases of TB patients, respectively. However, these figures translate into large absolute number of MDR-TB cases, who can transmit drug resistant strains to others, especially in overcrowding condition, high prevalence of HIV and malnutrition. For these reasons, this emerging problem calls for serious consideration and urgent action. (33)

According to the WHO global TB report 2011, there were an estimated 220,000 (261 per 100,000) incident cases of TB in Ethiopia in 2010. According to the same report the prevalence of TB was estimated to be 330,000 (394 per 100,000). There were an estimated 29,000 deaths (35 per 100,000) due to TB, excluding HIV related deaths, in Ethiopia during the same period (34).

As far as the TB control program is concerned, the country achieved 100% geographical and above 92% health facility DOTS coverage. Despite the extensive expansion of DOTS service in the country, the program performance indicators remain unsatisfactory, especially the case detection rate. In 2007/2008 the country achieved a case detection rate of smear positive TB of 34.5%, beside a treatment success rate of 84%. (33, 35)

Addis Ababa is the capital city of Ethiopia. Based on the 2007 census conducted by the Central Statistical Agency of Ethiopia (CSA), Addis Ababa has a population of about 3 million (36). Addis Ababa is composed by 10 sub-cities; each one composed by a certain number of Kebeles. Addis Ababa has 35 government health facilities and 40 private health facilities that offer anti-TB treatment. However, case detection for TB in Addis Ababa is 63 percent, lower than the WHO target of 70 percent. When patients are diagnosed, about 80 percent are successfully treated, mainly through the directly observed treatment short course (DOTS) (3).

3.8. Related Works

Tamer U. (5) conducted a research with title “Prediction the Existence of Mycobacterium Tuberculosis on Patients by Data Mining Approach”. The objective of the study was to develop a data mining solution which makes diagnosis of tuberculosis as accurate as possible and helps deciding if it is reasonable to start tuberculosis treatment on suspected patients without waiting the exact medical test results or not. The study was focused on classification of tuberculosis

patients. To make a correct diagnosis of tuberculosis, a medical test must be applied to patient's phlegm. The study has tried to applied different models on the data which are Bayesian Networks, Multilayer Perceptron, JRip (a RIPPER algorithm implementation), PART (a Partial Decision Trees algorithm implementation), RSES (a Rough Set algorithm implementation) and ANFIS (Adoptive Network based Fuzzy Inference system). The study was mostly focused on ANFIS method. The other methods that are used for comparison of ANFIS results. Based on the finding of their study, the researcher concluded that ANFIS is an accurate and reliable method comparing to Bayesian Network, Multilayer Perceptron, PART, JRip and RSES methods for classification of tuberculosis patients. (5)

Asha T., et al. (9) conducted a research with a title "Estimating the Statistical Significance of Classifiers used in the Prediction of Tuberculosis". The study was focused on classification of Tuberculosis with statistical significance is realized at two stages. First stage is the comparison of accuracies by classifying TB data into two categories Pulmonary Tuberculosis (PTB) and retroviral PTB (RPTB) i.e. TB along with AIDS using basic learning classifiers such as C4.5 Decision Tree, Support Vector Machines (SVM), K-nearest neighbor, Bagging and Naïve Bayesian algorithms. Second stage is evaluating the performance of these classifiers using paired test to select the optimum model. The results of the study showed that SVM is the best classifier, for their data set there is no much significant difference in the accuracies of SVM and Decision trees. SVM proves to be with 99.07 % accuracy against decision tree with 99%. Bagging proves to be better than K-nearest neighbor and lastly Naïve Bayes. The researchers conclude that the performance of a classifier in the classification of tuberculosis is dealt with statistical significance tests. It was found that SVM and Decision trees are not statistically significant whereas SVM and Naive Bayes have statistical significant differences between them. (9)

Asia Nesredin (11) conducted a research with a title "Mining Patients Data for Effective Tuberculosis Diagnosis" is used as a base to build the proposed study. The objective of study was to apply data mining techniques to patient's data to establish meaningful relationships or pattern for effective TB diagnosis. The research specifically looked at the use of clustering and classification data mining techniques. The study has tried to apply k-means clustering with some enhancements to aid in the process of segmenting the dataset into TB-positive and TB-negative. The resulting cluster is then used for developing the classification model. Classification was

employed in the study to identify patterns and predict the occurrence of TB. The classification task was made using J48 decision tree and Naïve Bayes classification algorithm and different experimentation was done. The model developed for predicting purpose has an accuracy of 85.93%. The discovered knowledge from the J48 decision tree is presented by traversing the tree for the easy of understand. (11)

Ali A. El-Solh, et al. (10) with a title "Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network". The objective of the study was to develop model with an artificial neural network using clinical and radiographic information to predict active pulmonary TB at the time of presentation at a health-care facility that is superior to physicians' opinion. The algorithm used by the study was a general regression neural network (GRNN) to develop the predictive model. Predictive accuracy was assessed by the c-index, which is equivalent to the area under the receiver operating characteristic curve. As a result neural network achieved a sensitivity of 100% (95% CI, 91 to 100%) and a specificity of 72% (95% CI, 65 to 77%). The physicians correctly diagnosed active pulmonary TB in 22 of 47 patients for a sensitivity of 47% (95% CI, 32 to 62%) and a specificity of 75% (95% CI, 71 to 79%). The corresponding c-indices (6 SEM) for the artificial neural network and the physicians were 0.947 6 +/- 0.028 and 0.61 6 +/- 0.045, respectively (p, 0.001). Finally the researchers conclude that an artificial neural network can identify patients with active pulmonary TB more accurately than physicians' clinical assessment. (10)

CHAPTER FOUR
RESEARCH METHODOLOGY,
DATA UNDERSTANDING AND PREPROCESSING

This chapter deals about research methodology, data understanding and preprocessing. The detail discussions about the project area, research design with its six steps, evaluating the creditability of the model, understanding of the data used for this study, descriptive statistical summary of selected attributes and finally discussion about preprocessing of the data employed through cleaning dataset and transforming continuous attributes to nominal.

4.1. Research Methodology

A research methodology is an arrangement of condition for collocation and analysis of data in a manner that aims to address the research problem. The research methodology used by this study was hybrid process model. This section deals with the project area, detailed discussion about research design with its six steps.

4.1.1. Project Area

The study area was in Addis Ababa city administration, which is the capital city of Ethiopia. With average elevation of 2500 meter above sea level, the city administration has a geographic and territorial possession with an area of 540 sq. km and a total population of about 3 million (36). There are 79 government owned health facilities in the city administration: 13 hospitals, 23 health centers, 9 clinics and 34 health posts. This study was conducted in two government hospitals namely, Menelik II hospital and St. Peter TB specialized hospital.

4.1.2. Research Design

This study is designed to predict the occurrence of TB based on the data collected from Menelik II and St. Peter hospitals. A hybrid data mining process model was applied. This process model was developed by adopting Knowledge Discovery in databases (KDD) process and Cross-industry Standard Process (CRISP-DM) models. Hybrid data mining model is selected for the present study, because:

- It provides more general, research-oriented description of the steps,

- It does emphasize the iterative aspects of the process, drawing experience from previous models, and
- It supports academia and industrial data mining projects.

Hybrid data mining process model has six steps: understanding of the problem, understanding the data, preparation of the data, data mining, evaluation and use of the discovered knowledge. The activities performed in each of the thus steps is presented below.

4.1.2.1. Understanding of the Problem

The researcher communicated with the subject experts and also read related literature on the problem domain to get familiar with the business, to get familiar with domain specific terminology, to determine the research objectives, to define problems and corresponding solutions, and to select the data mining tools appropriate to the problem domain.

4.1.2.2. Understanding of the Data

Understanding the data itself is a prerequisite for any data mining technique because, the result of a knowledge discovery process highly depends on the quality of data. In this phase, the original data collected for this research should be described briefly. This description includes listing attributes; identifying missing values and outliers, and evaluating the importance of attributes to the research goal.

To understand patient medical record, discussions were made with domain experts. In addition, to get familiar with the data and discover initial insights of the dataset exploratory data analysis was conducted using SPSS (statistical Package for Social Science) 16.0.

4.1.2.3. Preparation of the Data

In this phase data were undergone through data preparation to confirm the completeness of the data, to avoid redundancy, missing values, outliers and to verify the importance of attributes. The major tasks include: description of data sources, carrying out statistical summary measure, finding out distinct value, filling missing values, outlier and noisy data and data transformation / reduction activities were also undertaken in this phase so that the results meet the specific input requirements for DM tools.

4.1.2.4. Data Mining Techniques

In this phase different mining techniques were selected to attain the proposed research objectives. Predictive data mining techniques were used to develop the model capable of answering the stated problems. According to their popularity for the prediction of TB status, among the available algorithms in WEKA machine learning software, the researcher selected decision tree, naïve bayes, artificial neural network and support vector machine to build the models. Detailed discussion of the four mining algorithms is presented in section 2.4.

4.1.2.5. Evaluation of the Discovered Knowledge

The result of Knowledge discovery process was evaluated to reach at a certain conclusion which is relevant to the problem at hand. Evaluating the discovered knowledge also comprises understanding the results, checking whether the discovered knowledge is novel and interesting. The investigator used confusion matrix; accuracy, sensitivity, specificity and precision to evaluate the performance of each of the models.

4.1.2.6. Use of the Discovered Knowledge

This final phase consists of planning where and how to use the discovered knowledge. This last step determines the success of the entire knowledge discovery process. To facilitate the usability of the knowledge discovered, this study integrated a graphical user interface to support the prediction of TB status of patients. The results of this thesis work will be disseminated to the following stakeholders and to any interested parties. For this reason, interested domain experts and researchers can get access to the research results so as to support the decision making process, or use it for further research in the area or for any other applicable reasons.

- The result of the study will be presented to the school of information science and to the school of public health.
- A hardcopy of this thesis will be available at the bibliographic library of the school of information science.
- A softcopy of this thesis will be uploaded to Addis Ababa University electronic resource official website.

4.1.3. Evaluation of Credibility

The generated prediction model is evaluated to reach at a certain conclusion which is relevant to the problem at hand. It also comprises understanding the results, checking whether the discovered knowledge is novel and interesting. For classification, it is common to measure a classifiers performance in terms of the error rate or misclassification rate (17). The classifier predicts class label of each instances and if it is correct, counted as success, and if not counted as error (15).

There are a number of techniques used to measure accuracy such as the holdout (reserves a certain amount of data for testing and uses the remainders for training), random sub-sampling, k-fold cross-validation, and bootstrap method, which select datasets with replacement for training and testing purpose (15, 17).

This study used k-fold cross validation to partition the dataset in to training and test set, and then to measure the accuracy of the developed model. In cross-validation, you decide on a fixed number of folds, or partitions of the data. Then the data is split into K approximately equal partitions and each in turn is used for testing and the remainder is used for training. Once the number of folds is decided (partitions of the data) and the data is split in to K approximately equal partitions; each partitions in turn used for training and testing (17). This is called K-fold cross-validation.

This study used confusion matrix to evaluate the creditability of algorithms performance. Confusion matrix tells the correctly classified positive and negative records and incorrectly classified records are used as well. It is a tool for analyzing how well the classifier can recognize tuples of different classes (15). Given m classes, a confusion matrix is a table of at least m by m. For instance, in two class case with classes yes and no, reactive or non reactive, positive or negative, or lend or not lend, a single prediction has four different possible outcomes. The four possible outcomes of a two-class prediction are true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as shown in Table 4.1.

True positive (TP) and true negative (TN) are the correct classification. False positive (FP) occurs when the outcome is incorrectly predicted as yes when it is actually no. False negative (FN) are the positive tuples that were incorrectly predicted as no when it is actually yes (15, 17).

Table 4.1 The outcomes of two-class prediction

		Prediction Class		Total
		Positive	Negative	
Actual Class	Positive	TP	FN	TP + FN
	Negative	FP	TN	FP + TN
Total		TP + FP	FN + TN	TP + FP + FN + TN

Confusion matrix enable to assess how well the classifier can recognize “yes” tuples and how well it can recognize “no” tuples, the sensitivity (the true positive rate) and specificity (the true negative rate) measure can be used respectively (15).

4.1.4. Ethical Consideration

Since the study employed secondary data, the researcher did not have personal contact to the patients’ and did not need to ask consents. The research is governed by the following ethical considerations:

- The researcher did nothing with personal identifiers.
- The research is purely dedicated to academic purpose.
- Ethical clearance was obtained from ethical committee of St. Peters TB specialized hospital.
- The necessary verbal permission was obtained from Menelik II hospital director.
- The research is purely for public benefits.
- The research does not have harm on anybody in any ways.

4.1.5. Tools

The following tools have been used in the research work:

- Ms-Excel: - used for data preparation and pre-processing.
- WEKA 3.6: - used for building models, evaluation and analysis.
- SPSS 16.0: - used for data pre-processing and exploratory data analysis.
- Ms-Visual studio 2008: used for developing graphical user interface prototype.

4.2. Data Understanding

In order to meet the general objective of this research, collecting representative data is a prerequisite. Data understanding begins with collection of initial data. One of the main tasks in data mining process is analyzing and understanding the content and structure of the collected data.

4.2.1. Data Source and Data Collection

The data sources for this research are patient's real data (patient's medical records) taken from Menelik II and St. Peters TB specialized hospitals. This study used a total of 10,031 paper based patients' records, 5008 and 5023 from Menelik II and St. Peters TB specialized hospitals, respectively.

The patient's medical record contains all the necessary data for the research. The researcher and the data encoders hired by the researcher encoded all the data from medical records into Microsoft Excel workbook. This record consists of 15 attributes, considered as the most important and relevant information for the TB status prediction.

4.2.2. Description and Quality of Data

Series discussions were conducted with domain experts and review of related research on the domain area was made before selecting the target attributes from medical records. Because of these discussions and literature review, the researcher selected 15 attributes. Table 4.2 shows the attributes list along with their data description and data types. These attributes were selected due to their relevance or significance in diagnoses of tuberculosis.

Quality data is a prerequisite for any data mining tasks. To ensure data quality of this datasets; selection of data encoder was done by considering their experience in the research environment, orientation was given for data encoder about data quality and patient's privacy, and supervision were conducted during data collection by the primary investigator.

Table 4.2 Attributes description

No	Attributes	Meaning	Value	Data type
1	Sex	Sex of the patient	Male / Female	Nominal
2	Age	Age of the patient	Numeric age values of patient	Numeric
3	Weight	Weight of the patient	Numeric weight value of patient	Numeric
4	HIV Test result	Patient test result for HIV	Reactive, Non Reactive and Not Tested	Nominal
5	HIV Performed	Client tested for HIV	Yes/No	Nominal
6	Headache	Whether the patient have headache	Yes/No	Nominal
7	Cough	Whether the patient have cough for about 2 weeks	Yes/No	Nominal
8	Chest pain	Whether the patient have pain around the chest	Yes/No	Nominal
9	Bloody sputum	Sputum mixed with a blood	Yes/No	Nominal
10	Fever	An expected increase in temperature	Yes/No	Nominal
11	Weight loss	Whether the patient reduced in weight	Yes/No	Nominal
12	Night sweating	Whether the patient have sweats	Yes/No	Nominal
13	Loss of Appetite	Whether the patient have loss of appetite.	Yes/No	
14	Shortness of Breath	Whether the patient have problem in breathing	Yes/No	Nominal
15	TB Status	Tuberculosis test result of the patient	Positive / Negative	Nominal

4.3. Descriptive Statistical Summary of Selected attributes

The dataset has been described and visualized using Microsoft Excel and SPSS 16.0 to examine the properties of the dataset relative to the whole records. Simple statistical analysis has been performed to verify the quality of the dataset such as missing values, outliers and to obtain high level information regarding the data mining questions. Hence, the selected attributes are statistically described in details below. This is helpful for understanding of the dataset for experimentation.

4.3.1. Sex

The sex attribute describes the sex of the patient as Female and Male. Out of the total of 10031 patients, 5458 are females and 4573 are males as displayed in Figure 4.1.

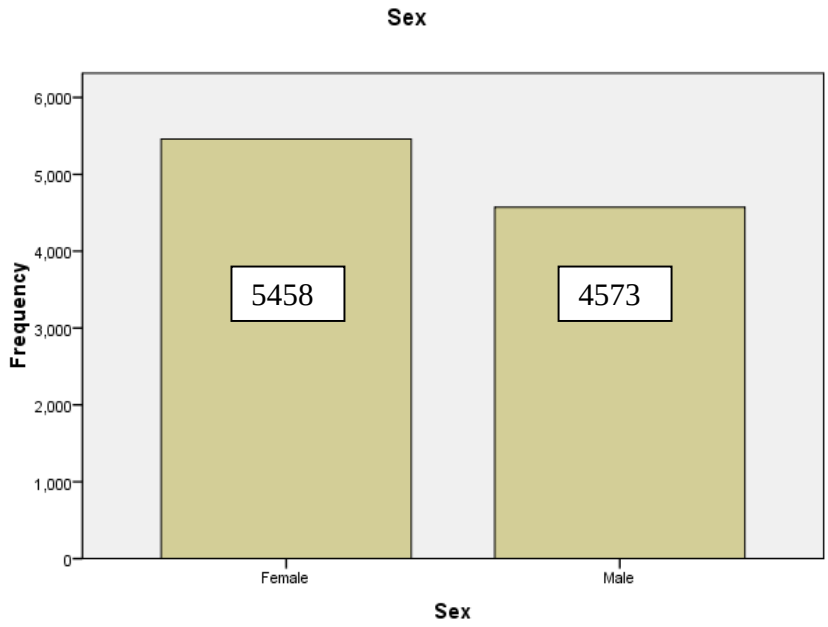


Figure 4.1 Sex distribution of patients diagnosed for TB

Table 4.3 shows the distribution of tuberculosis disease between male and female. The dataset shows that males are more likely to have the tuberculosis disease than females who took the test. Out of 5458 females who were diagnosed for TB, 49% were negative and the remaining (51%) were TB positive. Whereas, out of 4573 males who were diagnosed for TB, 45.9% were negative and the majority (54.1%) were positive

Table 4.3 TB status Cross tabulation by Sex

		TB status		
		Negative	Positive	Total
Sex	Female	2674 (49%)	2784 (51%)	5458
	Male	2099 (45.9%)	2474 (54.1)	4573
	Total	4773	5258	10031

4.3.2. Age

The age attribute contains the ages of the patients ranging from 0 year to 95 years. This attribute has been grouped into ranges of different year values (i.e. 0-14, 15-24, 25-49, 50-64, and above 64). Out of the total data, there are 3 missing values. Figure 4.2 shows that most of the patients diagnosed for TB are at the age range of 25–49.

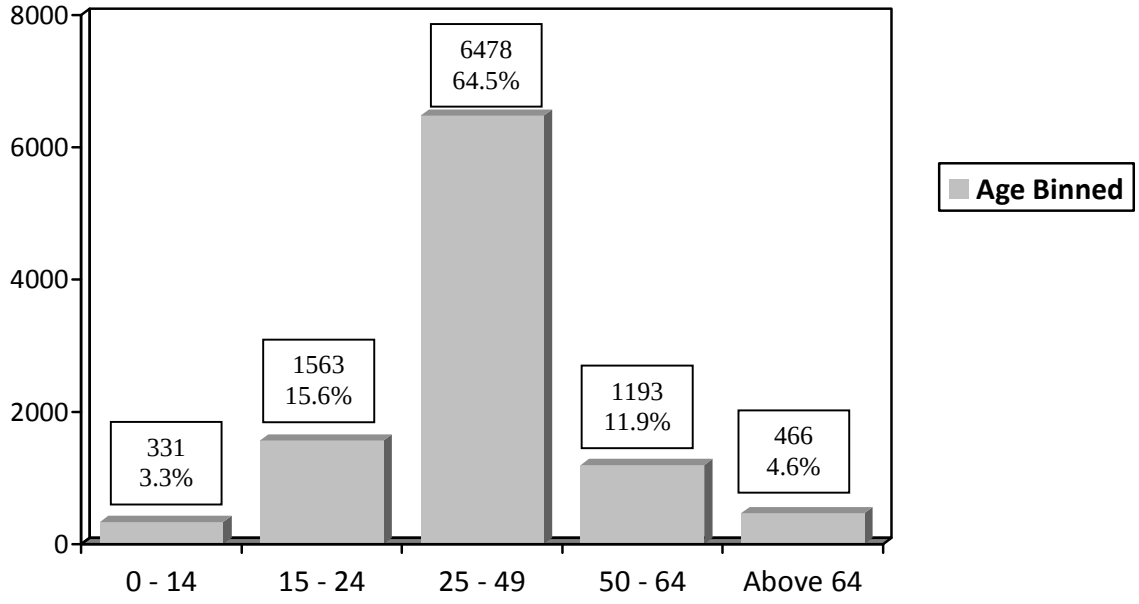


Figure 4.2 Age distribution of patients' diagnosed for TB

The distribution of TB disease within different age groups is displayed in Table 4.4. As shown in the Table, 64% of the TB patients are in the age range of 25 – 49, 17.4% of TB patients are in 15 – 24 age range, 11.7% of the patients are in 50 – 64 age range and the remaining 7.01% are patients in the age range of 0 – 14 and Above 64. Based on the above discussion, the researcher concludes that tuberculosis disease is more prevalent on young and productive age group.

Table 4.4 TB status cross tabulation by Age

		TB status		
		Negative	Positive	Total
Age (Binned)	1 – 14	239 (5%)	92 (1.75%)	331
	15 – 24	650 (14%)	913 (17.4%)	1563
	25 – 49	3102 (65%)	3376 (64.2%)	6478
	50 – 64	579 (12.1%)	614 (11.7%)	1193
	Above 64	204 (4.3%)	262 (5%)	466
Total		4773	5255	10028

4.3.3. HIV Test Result

This attribute indicates whether the patient has HIV or not. The attribute of HIV test result contains three different values such as reactive, non reactive and not tested. Figure 4.3 shows the HIV test results of patients.

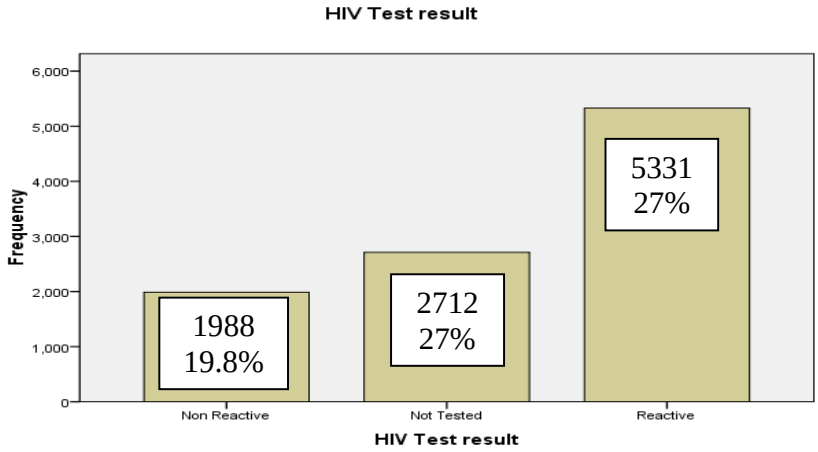


Figure 4.3 HIV test result of patients

4.3.4. Weight

This attribute indicates the weight of the patient. The weight of the patient in the data ranges from 6k.g to 108k.g. This attribute has not grouped into ranges of different values originally in the database. Weight attribute in this datasets contains 345 missing value because in some of patients medical records weight field are blank. In order to group the data in to different intervals, the researcher decided first to replace the missing values manually by average weight of each age and then grouped into different intervals by using visual binning method. Details of

binning process discussed in section 4.4.2. The statistical summary of the weight attribute is displayed in Figure 4.4.

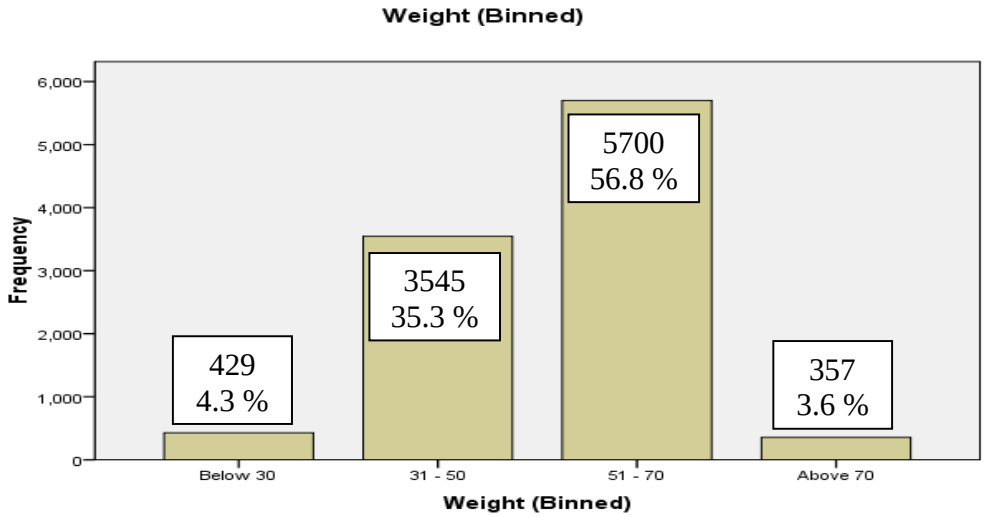


Figure 4.4 Distribution of patients' weight

The distribution of tuberculosis disease with in different weight group is shown in Table 4.5. Based on this dataset, most TB positive patients are grouped between the body weight of 31 and 50.

Table 4.5 TB status Cross tabulation by Weight

		TB status		
		Negative	Positive	Total
Weight (Binned)	Below 30	247 (57.5%)	182 (42.4%)	429
	31 – 50	1282 (36.2%)	2263 (63.8%)	3545
	51 – 70	3020 (53%)	2680 (47%)	5700
	Above 70	224 (62.7%)	133(37.3%)	357
Total		4773	5258	10031

4.3.5. TB Status

This attribute indicates whether the patients have TB disease or not. It has two possible values, namely positive or negative. Table 4.6 shows, out of 10031 records, 4773 patients were TB negative and the remaining 5258 patients were TB positive.

Table 4.6 Summary of Patients' TB status

TB Status	Frequency	Percent	Cumulative Percent
Negative	4773	47.6	47.6
Positive	5258	52.4	100.0
Total	10,031	100.0	

4.3.6. HIV Performed

This attribute indicates whether the patient tested for HIV or not. The attribute of HIV performed contains two different values such as yes and no. Figure 4.5 shows the statistical summary of HIV performed attribute.

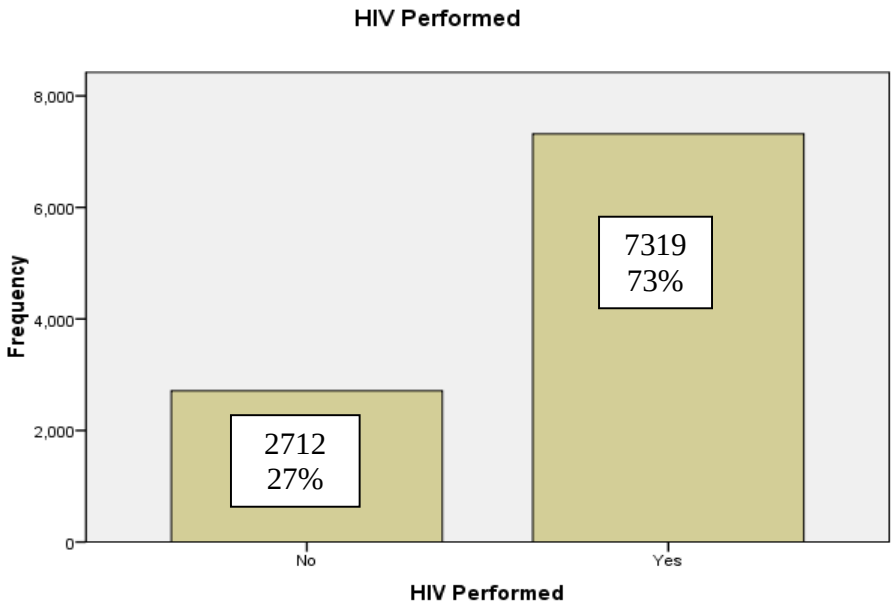


Figure 4.5 HIV performed of patients'

4.3.7. Clinical Symptoms of Patients'

The clinical symptom of patients' such as headache, cough, chest pain, bloody sputum, fever, weight loss, night sweating, loss of appetite and shortness of breath were contains two possible

values such as yes and no. The statistical described of patients' clinical symptom presented in Table 4.7.

Table 4.7 Summary of patients' clinical symptoms

Clinical Symptom	Frequency and Percent	
	Yes	No
Headache	3297 (32.9%)	6734 (67.1%)
Cough	8858 (88.3%)	1173 (11.7%)
Chest Pain	5831 (58.1%)	4200 (41.9%)
Bloody Sputum	640 (6.4%)	9391 (93.6%)
Fever	5314 (53%)	4717 (47%)
Weight Loss	3635 (36.2%)	6396 (63.8%)
Night Sweating	6084 (60%)	3947 (39.3%)
Loss of Appetite	4229 (42.2%)	5802 (57.8%)
Shortness of Breath	3357 (33.5%)	6674 (66.5%)

4.4. Data Pre-processing

This step consists of the core of data mining and takes much of the research time and effort of the entire knowledge discovery process but, is the most necessary activity in data mining process (15). Data preprocessing involves all the action taken before the actual data analysis process starts. There are a number of data preprocessing tasks involved in this study such as data cleaning, data transformation and data reduction techniques. Each of these techniques are discussed in detail as follows.

4.4.1. Data Cleaning

Data cleaning is a time-consuming and labor-intensive procedure but it is absolutely necessary for successful data mining (17). Many of the existing data are not clean; they need the effort of the researcher in the area to get clear of unclean data. Data cleaning is a process which involves filling in missing values, smooth noisy data, identify and remove outliers, and resolving inconsistencies.

4.4.1.1. Handling Missing Values

Missing values are frequently indicated by out of range entries, perhaps a negative number in a numeric field that is normally only positive or a 0 in a numeric field that can never normally be 0. For nominal attributes, missing values may be indicated by blanks or dashes. Most datasets encountered in practice contain missing values (17). Identifying and solving those missed values is needed to prepare the dataset for experimentation. Because of their popularity and familiarity of the researcher with these software, the researcher has used Microsoft Excel and SPSS 16.0 for exploratory analysis, identified missing values and took necessary measure to handle missing values.

In this dataset, two attributes namely age and weight consist 3 and 345 missing values, respectively. For handling those missing values, the recommended technique is replacing the missing value with the attribute mean for numeric variable; this method has been used to replace missed value for Age attribute. But replacing missing value of patient weight with mean value for all age group patients is not convenient because people with different age have different weight in average, so the researcher decided using manual technique to replace missing value of weight attribute with average weight of each age.

4.4.2. Data Transformation

Data transformation is about transforming or consolidating the data to make it appropriate for mining. The process of data transformation can include smoothing (includes binning, regression, and clustering which works to remove noise from the data), generalization (where low-level raw data are replaced by higher-level concepts. For example numeric attribute of age may be generalized to youth, middle age, and so on) and normalization (It is an operation performed on attribute to scale the value that fall within a small specified range) (15).

Among those techniques of transformation, data discretization / binning is selected by the researcher. Binning is used to reduce data size by dividing the range of a continuous attribute in to interval. Interval labels can then be used to replace the actual data values (17). Divide the range of continuous values into N intervals of equal size and give label for each interval. The transformation of weight attribute was done through observation of the dataset and discussion with domain expert. Table 4.8 shows the result of binned weight attribute and their label.

Table 4.8 Discretized result of patients' weight

Weight (Binned) Label	Frequency	Percent	Cumulative Percent
Below 30	429	4.3	4.3
31 – 50	3545	35.3	39.6
51 – 70	5700	56.8	96.4
Above 70	357	3.6	100.0
Total	10031	100.0	

The other transformation done in this research is transforming the numeric value of age attribute to nominal. In order to transform the value of age attribute, the researcher decided to categorize patients' age in to 5 width age ranges. Table 4.9 shows the binned age of patient.

Table 4.9 Discretized result of Patients' age

Age (Binned)	Frequency	Percent	Cumulative Percent
1 – 14	331	3.3	3.3
15 – 24	1563	15.6	18.9
25 – 49	6478	64.5	83.5
50 – 64	1193	11.9	95.4
Above 64	466	4.6	100.0
Total	10031	100.0	

CHAPTER FIVE
DATA MINING AND MODEL SELECTION

In this chapter, the researcher describes the techniques that have been used in developing models to predict TB status of patients. Data mining classification methods were chosen to develop predictive models. Different experiments were done using four algorithms: Decision Tree, Naïve Bayes, Support Vector Machine and Artificial Neural Network in WEKA 3.6.8. This chapter gives details of the experiments and performance of the models.

5.1. Experimental Setup

Datasets for this research were first encoded to Microsoft excel worksheet in XLS file format and then converted to comma-separated value (CSV) file format (as it can be seen in appendix 1) because dataset with CSV file formats can be loaded to WEKA software. XLS is default file format for spread sheet program. But WEKA data mining software by default store data as an attribute-relation file format (ARFF), which means the datasets are needed to be converted from a spreadsheet to ARFF. There are two ways of representing datasets (that consist of independent, unordered instances and not involved relationships among instances) in an ARFF file format (17).

Mostly, spreadsheet and database programs allow exporting data into a file in CSV file format as a list of records with commas in between. Then you only load the file into text editor or word processor and add the dataset's name using the @relation tag, the attribute information using @attribute tag, and data information with @data tag and save the file as raw text with ARFF file format (17). Another method used to create ARFF file is to simply load the dataset in CSV file format to WEKA software, and click on the save button (in the preprocess tab) and save the dataset with ARFF file format. Figure 5.1 shows the sample of machine understandable format of the dataset in WEKA employed for this study. Dataset in ARFF file format was converted from CSV by WEKA 3.6.8 software during data preprocessing.

In this study 10,031 instances are ready for experimentation after data preprocessing. In the first chapter of this document the researcher planned to compare the performance of the four

classification algorithms as dataset size increases, three datasets were created by partitioning the whole dataset:

- Dataset 1: consists of 5008 records from Menelik II hospital.
- Dataset 2: consists of 7508 records (5008 from Menelik II hospital and 2500 from St. Peters TB specialized hospital).
- Dataset 3: consists of 10,031 records (5005 from Menelik II hospital and 5023 from St. Peters TB specialized hospital).

Note: Model comparison and best model selection was done using only dataset 3 (the main dataset). Dataset 1 and 2 were created only for the purpose of comparing the performance of mining algorithms as the dataset size increases.

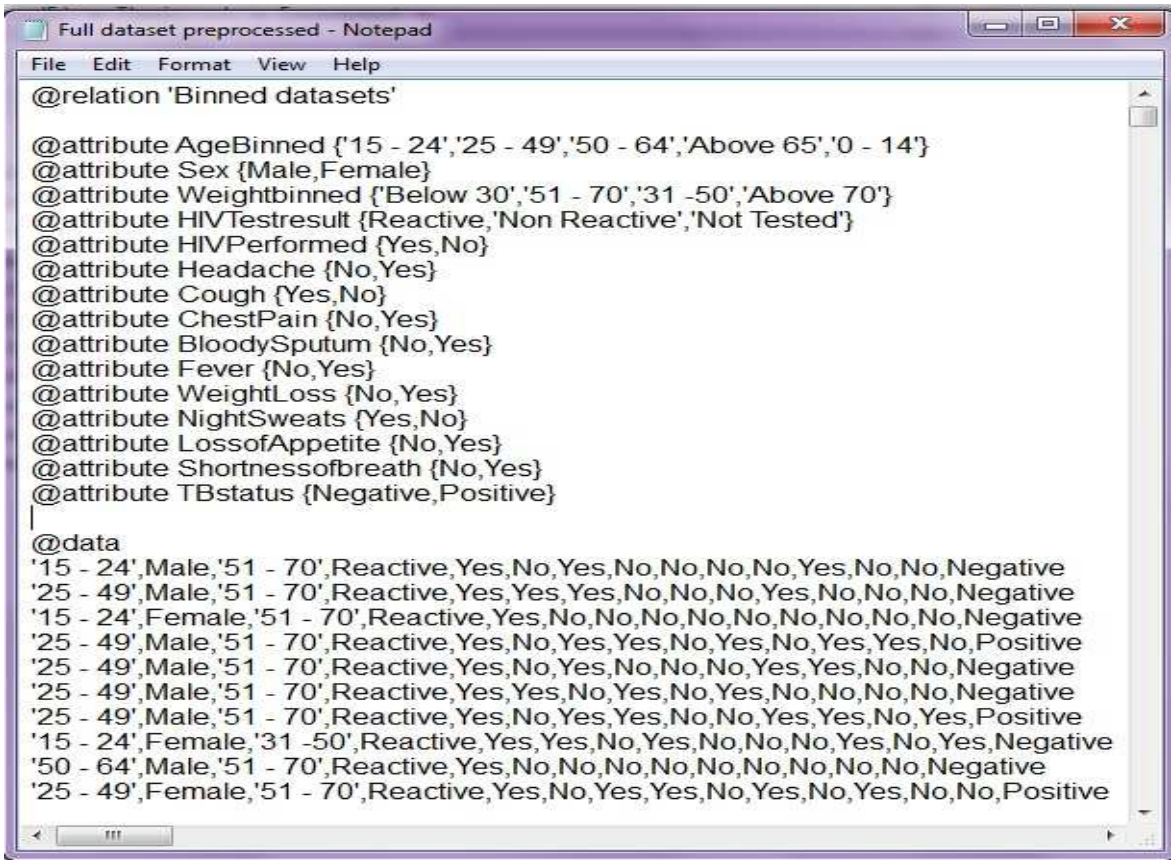


Figure 5.1 ARFF file of the dataset generated by WEKA.

The modeling phase in the data mining process of this investigation was carried out in three phases and two sub-phases. The three phases of experimentation were conducted using the three datasets. The two sub phases were carried out in this study by:

- Using all the 15 attributes of the datasets
- Using 7 best attributes of the datasets selected by the data mining attribute selection algorithm.

In this research, eight experiments were conducted with the following experiment design of four data mining algorithms and attribute selection. Each of the experiments is repeated three times using the three datasets (once for each dataset).

1. Experiment I: using J48 decisions tree with default parameters and applying 15 attributes.
2. Experiment II: using J48 decision tree with default parameters and 8 best selected attributes.
3. Experiment III: using naïve bayes with default parameters and applying 15 attributes.
4. Experiment IV: using naïve bayes with default parameters and applying 8 best selected attributes.
5. Experiment V: using Sequential Minimal Optimization algorithm of SVM with default parameters and applying 15 attributes.
6. Experiment VI: using Sequential Minimal Optimization algorithm of SVM with default parameters and applying 8 best selected attributes to the three datasets.
7. Experiment V: using Multilayer perceptron of ANN with default parameters and applying 15 attributes.
8. Experiment VI: using Multilayer perceptron of ANN with default parameters and applying 8 best selected attributes.

The test option used by this study was 10 fold cross validation for partition of the datasets into training and test set. Data was divided into 10 folds, and each in turn is used for testing and the remainder is used for training. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning procedure is executed a total of 10 times on different training sets (17). The researcher decided to

use 10 folds because extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up (17). Before presenting the details of each of the experiments conducted in this study, it is reasonable to describe first how the best attributes are selected.

5.2. Effect of Replacing Missing Values

In this datasets, two attributes namely age and weight consist 3 and 345 missing values, respectively. Records with missing values are about 3.7% of the total records. These missing values were replaced by average mean for age and average weight of each age for weight. The researcher decides to conduct experiments to identify the effect of replacing missing values. These experiments conducted using J48 decision tree classifier applied to dataset with replacing missing values and dataset with removed records with missing values. As shown in the Table 5.1, the performance of J48 decision tree was better using dataset with replacing missing values than dataset with removing records.

Table 5.1 Effect of replacing missing values

Model	Instance Number	No of Leaves	Size of Tree	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
J48 with all attributes was applied to dataset with removing records.	9683	82	150	0.951	0.969	0.937	0.964	0.049	95.01%
J48 with all attributes was applied to dataset with replaced missing values.	10,031	98	177	0.953	0.972	0.942	0.964	0.047	95.24%

Therefore, the researcher decides to using dataset with replaced missing values for further experimentation.

5.3. Attribute Selection

To proceed with the experiments, feature subset selection is performed. Feature selection is the process of removing features from the dataset that are irrelevant with respect to the task that is to be performed. Attribute selection is useful for reducing the dimensionality of the data by deleting unsuitable attributes which improves the performance of learning algorithms. It also speeds them up, although this may be outweighed by the computation involved in attribute selection. More importantly, dimensionality reduction yields a more compact, more easily interpretable representation of the target concept, focusing the user's attention on the most relevant variables. In general, feature selection techniques can be categorized into two: filter methods and wrapper methods. Filter methods operate independently of the learning algorithms while wrapper methods take into account the learning algorithms to be used (17). To select the best attribute subset selector, the investigator compared two attribute subset selectors: CFS (correlation-based feature subset selection) subset evaluator with best first search method and information gain attribute evaluator with ranker search method.

As it is seen in appendix 2, CFS subset evaluator select best attribute by evaluating the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. CFS subset evaluator selected 7 best attributes, such as HIV test result, cough, Chest Pain, Bloody sputum, Weight loss, Night sweats and Shortness of breath.

Information gain attribute evaluator works by evaluating the worth of an attribute by measuring the information gain with respect to the class. As can be seen in appendix 3, information gain attribute subset evaluator algorithm ranks the attributes based on the information gain with respect to class. The researcher selected 7 best attributes according to their rank from 14 independent attributes. Table 5.2 presented the best 7 attributes that were selected based on rank given by information gain attribute evaluator algorithm.

Table 5.2 Best attributes by information gain evaluator.

Rank	Attribute Name	Possible Value	Data Type
1	Shortness of breath	Yes / No	Nominal
2	Chest Pain	Yes / No	Nominal
3	Cough	Yes / No	Nominal
4	Weight loss	Yes / No	Nominal
5	Loss of appetite	Yes / No	Nominal
6	Night sweats	Yes / No	Nominal
7	HIV test result	Reactive / Non reactive / Not tested	Nominal

The comparison between two attribute subset selections was done based on accuracy, specificity and sensitivity of a model developed using J48 classifier with best 8 attributes (7 attributes selected by CFS subset evaluator plus the class attribute-TB status and 7 attributes selected by information gain attribute evaluator plus the class attribute). The result in Table 5.3 shows that the model generated by J48 classifier with 8 best attributes selected by information gain attribute evaluator performed better than the model generated by 8 best attributes selected by CFS attribute subset evaluator.

Table 5.3 Comparison of two attribute subset selector.

J48 with 8 best attributes applied to dataset 3 (main dataset)	Selection by CFS subset evaluator			Selection by information gain attribute evaluator		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
	93.3%	0.936	0.929	94.45%	0.933	0.957

Therefore, the best attributes selected by information gain attribute evaluator are used for further experimentation.

5.3. Decision Tree Classifiers

Experiment 1

In this experiment, the performance of J48 classifier in predicting TB status of patients using the three datasets was evaluated. Three models were built using J48 decisions tree with default parameters and all (15) attributes.

Table 5.4 Decision tree result with all attributes.

Model	Instance Number	No of Leaves	Size of Tree	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
J48 with all attributes was applied to dataset 1	5008	65	115	0.07	0.96	0.974	0.964	0.956	0.041	96.1%
J48 with all attributes was applied to dataset 2	7508	82	145	0.21	0.953	0.966	0.944	0.963	0.045	95.29%
J48 with all attributes was applied to dataset 3	10,031	98	177	0.6	0.953	0.972	0.942	0.964	0.047	95.24%

Table 5.4 shows the results (based on ten-fold cross validation) of experiment 1. With dataset 1, out of 5008 records, 4813 were correctly classified and the remaining 195 (4%) instances were incorrectly classified and, precision and receiver operating character curve (ROC area) of the model were 0.96 and 0.974 respectively. With dataset 2, out of 7508 records 7154 were correctly classified and the remaining 351 (4.6%) instances were incorrectly classified and, precision and ROC area of the model were 0.953 and 0.966 respectively. With dataset 3, out of 10,031 records 9554 were correctly classified and the remaining 477 (4.7%) instances were incorrectly classified and, precision and ROC area of the model were 0.953 and 0.972 respectively.

Experiment 2

The second experiment was conducted to investigate the performance of J48 classifier with the 8 best selected attributes. Three models were built using J48 decisions tree with default parameters and best 8 attributes using the three datasets. There are three scenarios in this experiment for testing the performance of J48 classifier.

Table 5.5 shows the results of experiment 2. With dataset 1, out of 5008 records 4732 were correctly classified and the remaining 276 (5.5%) instances were incorrectly classified and, precision and ROC area of the model were 0.945 and 0.974 respectively. With dataset 2, out of 7508 records 7085 were correctly classified and the remaining 423 (5.6%) instances were incorrectly classified and, precision and ROC area of the model were 0.944 and 0.962 respectively. With dataset 3, out of 10,031 records 9474 were correctly classified and the remaining 557 (5.5%) instances were incorrectly classified and, precision and ROC area of the model were 0.945 and 0.968 respectively.

Table 5.5 Decision tree Results with best attributes

Model	Instance Number	No of Leaves	Size of Tree	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
J48 with 8 best attributes was applied to dataset 1	5008	25	48	0.05	0.945	0.974	0.95	0.936	0.041	94.48 %
J48 with 8 best attributes was applied to dataset 2	7508	30	61	0.06	0.944	0.962	0.942	0.946	0.052	94.4 %
J48 with 8 best attributes was applied to dataset 3	10,031	35	66	0.5	0.945	0.968	0.933	0.957	0.054	94.44 %

5.3.1. Decision Tree Evaluation

Six experiments were conducted based on all 15 attributes and 8 best selected attributes within three datasets to evaluate the performance of J48 classifier. According to this datasets, the results

of J48 classifier in Table 5.4 and Table 5.5 show that when the size of dataset increased, the number of leaves, size of tree, false prediction (FP) rate and time taken to build model were increased, whereas true prediction rate, precision, ROC area and accuracy of the model were decreased.

The attribute subset selection did not improve the performance of J48 classifier according to the results of experiment 1 and 2. The number of leaves, size of tree, true positive (TP) rate (sensitivity), true negative (TN) rate (specificity), precision, ROC area and accuracy of the generated models were decreased because of attribute subset selection, whereas false prediction rate of the models were increased.

5.4. Naïve Bayes Classifiers

Experiment 3

The purpose of this experiment was to evaluate the performance of Naïve Bayes classifier algorithm using the three datasets. Three models were built using naïve bayes classifier with default parameters and 15 attributes.

Table 5.6 Naïve Bayes result with all attributes.

Model	Instance Number	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
Naïve Bayes with all attributes was applied to dataset 1	5008	0.01	0.902	0.961	0.883	0.92	0.094	89.63 %
Naïve Bayes with all attributes was applied to dataset 2	7508	0.01	0.872	0.951	0.844	0.913	0.116	87.2%
Naïve Bayes with all attributes was applied to dataset 3	10,031	0.01	0.877	0.956	0.834	0.916	0.123	87.3%

Table 5.6 presents the results of experiment 3. With dataset 1, out of 5008 instances 4489 were correctly and the remaining 519 (10.36%) were incorrectly classified and, the precision and ROC area of the model were 0.902 and 0.961 respectively. With dataset 2, out of 7508 instances 6546 were correctly and the remaining 962 (12.8) were incorrectly classified and, precision and ROC area of the model were 0.872 and 0.951 respectively. With dataset 3, 8757 instances were

correctly classified and 1274 (12.7) instances were incorrectly classified and, the precision and ROC area of the model were 0.877 and 0.956 respectively.

Experiment 4

Three models were built using naïve bayes classifier with default parameters and 8 best selected attributes applied to the three datasets. The fourth experiment has been tested with three scenarios using 10-fold cross validation.

Table 5.7 presents the results of experiment 4. In scenario 1, the result shows that 4497 instances were correctly classified and the remaining 511 (10.2%) instances were incorrectly classified out of 5008 instances and, the precision and ROC area of the model were 0.902 and 0.964 respectively. In scenario 2, out of 7508 instances 6573 were correctly classified and the remaining 935 (12.45%) were incorrectly classified and, precision and ROC area of the model were 0.88 and 0.954 respectively. In the third scenario, 8957 instances were correctly classified and 1074 (10.70%) instances were incorrectly classified out of 10,031 instances and, the precision and ROC area of the model were 0.898 and 0.96 respectively.

Table 5.7 Naïve Bayes results with best attributes

Model	Instance Number	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
Naïve Bayes with 8 best attributes was applied to dataset 1	5008	0.01	0.902	0.964	0.877	0.933	0.088	89.79 %
Naïve Bayes with 8 best attributes was applied to dataset 2	7508	0.01	0.88	0.954	0.883	0.954	0.079	87.54 %
Naïve Bayes with 8 best attributes was applied to dataset 3	10,031	0.01	0.898	0.96	0.849	0.941	0.103	89.29 %

5.4.1. Naïve Bayes Evaluation

Six experiments were conducted based on all 15 attributes and 8 best selected attributes within three datasets to evaluate the performance of naïve bayes classifier. According to these datasets,

the results in Table 5.6 and Table 5.7 show that naïve bayes classifier generate inconsistent results when the dataset size increased.

However, the attribute subset selection did improve the performance of in naïve bayes classifier according to the result of experiment 3 and 4. The experiments conducted with 8 selected attributes have high model accuracy than the corresponding experiments conducted by all attributes within the same datasets.

5.5. Support Vector Machine based classifiers

Experiment 5

Experiment five was conducted to evaluate the performance of sequential minimal optimization (SMO) classifier algorithm to predict TB status of patients within the three datasets. Three models were built using SMO with default parameters and all attributes were applied to the three datasets. Experiment 5 has been tested with three scenarios using 10-fold cross validation.

Table 5.8 Support vector machine result with all attributes.

Model	Instance Number	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
SMO with all attributes was applied to dataset 1	5008	8.94	0.952	0.95	0.956	0.943	0.052	95.12%
SMO with all attributes was applied to dataset 2	7508	26.03	0.947	0.947	0.952	0.928	0.063	94.12%
SMO with all attributes was applied to dataset 3	10,031	28.4	0.939	0.939	0.945	0.932	0.062	93.89%

Table 5.8 presents the results of experiment 5. In scenario 1, the result shows that out of 5008 instances, 4764 were correctly classified and the remaining 244 (4.87%) were incorrectly classified and, the precision and ROC area of the model were 0.952 and 0.95 respectively. In scenario 2, the result shows that out of 7508 instances, 7071 were correctly classified and the remaining 437 (5.8%) were incorrectly classified and, the precision and ROC area of the model were 0.947 and 0.947 respectively. In scenario 3, the result shows that out of 10,031 instances,

9419 were correctly classified and the remaining 612 (6.1%) were incorrectly classified, and the precision and ROC area of the model were 0.939 and 0.939 respectively.

Experiment 6

Three models were built using SMO classifier with default parameters and best 8 selected attributes applied to the three datasets. Experiment 6 has been tested with three scenarios using 10-fold cross validation.

Table 5.9 Support vector machine results with best attributes

Model	Instance Number	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
SMO with 8 best attributes was applied to dataset 1	5008	3.1	0.94	0.937	0.951	0.923	0.067	94.03%
SMO with 8 best attributes was applied to dataset 2	7508	7.42	0.932	0.932	0.9	0.938	0.077	91.58%
SMO with 8 best attributes was applied to dataset 3	10,031	9.79	0.917	0.917	0.903	0.931	0.083	91.59%

Table 5.9 presents the results of experiment 6. In scenario 1, the result shows that 4709 were correctly classified and the remaining 299 (5.9%) were incorrectly classified out of 5008 instances, and the precision and ROC area of the model were 0.94 and 0.937 respectively. In scenario 2, the experiment result shows that out of 7508 instances, 6876 were correctly classified and the remaining 632 (8.4%) were incorrectly classified and, the precision and ROC area of the model were 0.932 and 0.932 respectively. In scenario 3, the experiment result shows that out of 10,031 instances, 9188 were correctly classified and the remaining 843 (8.4%) were incorrectly classified and, the precision and ROC area of the model were 0.917 and 0.917 respectively.

5.5.1. SMO Evaluation

Six experiments were conducted based on all 15 attribute and 8 best selected attributes within three datasets to evaluate the performance of SMO classifier. According to this datasets, the

results of SMO classifier in Table 5.8 and Table 5.9 show that when the dataset size increased the true positive rate, true negative rate, precision, ROC area and accuracy of the model decreased, whereas false prediction rate and time taken to build model were increased.

The attribute subset selection did not improve the performance of SMO classifier according to the results of experiment 5 and 6. The true prediction rate, precision, ROC area and accuracy of the generated model were decreased, whereas false prediction rate was increased due to attribute subset selection.

5.6. Artificial Neural Network based classifiers

Experiment 7

Three models were built using multilayer perceptron (MLP) classifier with default parameters and all attributes applied to the three datasets. Experiment 7 has been tested with three scenarios using 10-fold cross validation.

Table 5.10 MLP result with all attributes.

Model	Instance Number	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
MLP with all attributes was applied to dataset 1	5008	31.94	0.962	0.987	0.968	0.95	0.04	96.1%
MLP with all attributes was applied to dataset 2	7508	48.17	0.953	0.978	0.957	0.948	0.04	95.29%
MLP with all attributes was applied to dataset 3	10,031	61.37	0.952	0.978	0.948	0.958	0.047	95.23%

Table 5.10 shows the results of experiment 7. With dataset I, out of 5008 instances 4816 were correctly classified and the remaining 192 (3.8%) were incorrectly classified and, the precision and ROC area of the model were 0.962 and 0.987 respectively. With dataset II, out of 7508 instances 7155 were correctly classified and the remaining 353 (4.7%) were incorrectly classified and, the precision and ROC area of the model were 0.953 and 0.978 respectively. With dataset II, out of 10,031 instances 9553 were correctly classified and the remaining 478 (4.76%) were

incorrectly classified and, the precision and ROC area of the model were 0.952 and 0.978 respectively.

Experiment 8

Three models were built using MLP classifier with default parameters and best 8 selected attributes applied to the three datasets. Experiment 8 has been tested with three scenarios using 10-fold cross validation.

Table 5.11 MLP result with best attributes.

Model	Instance Number	Time(sec)	precision	ROC area	TP Rate	TN Rate	FP Rate	Accuracy
MLP with 8 best attributes was applied to dataset 1	5008	8.86	0.951	0.983	0.948	0.953	0.049	94.98%
MLP with 8 best attributes was applied to dataset 2	7508	15.39	0.95	0.981	0.94	0.944	0.051	94.23%
MLP with 8 best attributes was applied to dataset 3	10,031	17.22	0.945	0.975	0.936	0.955	0.054	94.46%

Table 5.11 shows the results of experiment 10. In scenario 1, the result shows that out of 5008 instances, 4757 were correctly classified and the remaining 251 (5.01%) were incorrectly classified, and the precision and ROC area of the model were 0.951 and 0.983 respectively. In scenario 2, the result shows that out of 7508 instances, 7077 were correctly classified and the remaining 431 (5.7%) were incorrectly classified, and the precision and ROC area of the model were 0.95 and 0.981 respectively. In scenario 3, the result shows that out of 10,031 instances, 9476 were correctly classified and the remaining 555 (5.5%) were incorrectly classified, and the precision and ROC area of the model were 0.945 and 0.975 respectively.

5.6.1. Multilayer Perceptron Evaluation

Six experiments were conducted based on all 15 attributes and 8 best selected attributes within three datasets to evaluate the performance of MLP classifier. According to these datasets, the results of MLP classifier in Table 5.10 and Table 5.11 show that when the dataset size increased,

the true positive rate, precision, ROC area and accuracy of the model were decreased, whereas false prediction rate and time taken to build model were increased.

The attribute subset selection did not improve the performance of MLP classifier according to the results of experiment 7 and 8 results. The true prediction rate, precision, ROC area and accuracy of the generated model were decreased, whereas false prediction rate was increased due to attribute subset selection.

5.7. Discussion

To develop predictive models, J48, Naïve Bayes, SMO and MLP algorithms were used. A total of 24 experiments were conducted based on all attributes and selected 8 attribute within the three datasets. The experiments were designed to investigate to whether increasing dataset size improves or degrades the performance of the four algorithms; to evaluate the effect of attribute subset selection on the performance of the four algorithms and to compare the performance of the algorithms in predicting TB status of patients.

5.7.1. Effect of Attribute subset selection

The effect of attribute subset selection did improve the performance of naïve bayes classifier. The experiments conducted by naïve bayes classifier with 8 selected attributes have high model accuracy than the corresponding experiments conducted by all attributes within the same datasets.

However, the attribute subset selection did not improve the performance of J48, SMO and MLP classifiers. Due to attribute subset selection, the true prediction rate, precision, ROC area, correctly classified instances and accuracy of the models were decreased, whereas false prediction rate of the models was increased compared with the results of the experiments conducted using all attributes. Figure 5.2 shows the accuracy of the classifiers before and after attribute subset selection using dataset 3.

As shown in Figure 5.2 the performance of J48, SMO and MLP classifier decreased due to attribute subset selection. This study conducted additional experiment to examine why the attribute subset selection did not improve the performance of these classifiers.

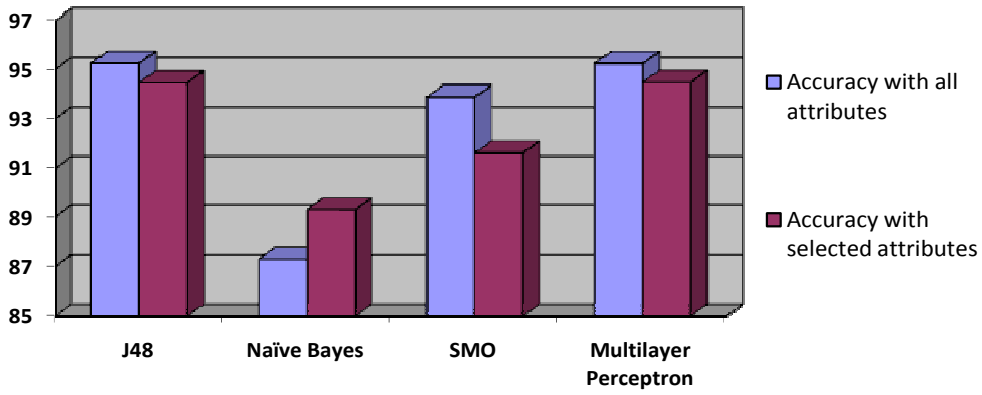


Figure 5.2 Effect of attribute selection on classification accuracy.

Experiments were conducted to get the reason why the performance of the algorithms decline with the 8 best selected attributes. These experiments were conducted to search for the attribute/s that is/are not chosen as best attribute/s by information gain attribute evaluator but boost/s the performance of the classifier. The newly added attribute on the selected 8 attributes improve or deteriorate the performance of the models compared to the result of experiment conducted with 8 best attributes. However, the performances of the classifiers with 9 attributes were still poor compared to the performances of a classifier with all attributes.

Table 5.12 presents the effect of attribute subset selection in the performance of models with 8 and 9 attributes as compared with all attributes. The results in the table show that the best performing model was J48 classifier with all attributes. Some of the experiments conducted using 9 attributes performed better than 8 best attributes, whereas others performed worse than the model developed using 8 best attributes.

Table 5.12 Searching best attribute boost performance of J48 classifier.

Model	Time (sec)	TP Rate	TN Rate	FP Rate	Precision	ROC area	Accuracy
J48 with all attributes	0.6	**0.942**	**0.964**	**0.047**	**0.953**	**0.972**	**95.24%**
J48 with 8 selected attributes	0.5	0.933	0.957	0.054	0.945	0.968	94.44%
J48 with 8 selected attributes plus Bloody sputum	0.83	0.941	0.957	0.051	0.949	0.97	94.85%
J48 with 8 selected attributes plus Headache	0.2	0.933	0.961	0.052	0.947	0.968	94.61%
J48 with 8 selected attributes plus Fever	**0.13**	0.934	0.96	0.053	0.947	0.968	94.16%
J48 with 8 selected attributes plus Sex	**0.13**	0.933	0.956	0.055	0.944	0.968	94.38%
J48 with 8 selected attributes plus weight	0.17	0.933	0.956	0.055	0.944	0.967	94.37%
J48 with 8 selected attributes plus Age	0.14	0.933	0.957	0.054	0.945	0.967	94.43%

As shown in the Table 5.12, the experiments conducted using 9 attributes (best 8 selected plus Bloody sputum or Headache) perform better than 8 best attributes but still poor compared with all attributes. Therefore, the researcher decided to conduct additional experiment with 10 attributes (best 8 selected plus Bloody sputum and Headache). However, the performance of the classifiers with 10 attributes was still poor compared to the performances of a classifier with all attributes.

Table 5.13 presents the effect of attribute subset selection in the performance of the models with 8 best, 9 and 10 attributes as compared with all attributes. The results in the table show that the best performing model was J48 classifier with all attributes. The experiment conducted using 10 attributes perform better than 8 best attributes and 9 attributes.

Table 5.13 Effect of attributes in the performance of J48 classifier

Model	Time (sec)	TP Rate	TN Rate	FP Rate	Precision	ROC area	Accuracy
J48 with all attributes	0.6	**0.942**	**0.964**	**0.047**	**0.953**	**0.972**	**95.24%**
J48 with 8 selected attributes	0.5	0.933	0.957	0.054	0.945	0.968	94.44%
J48 with 8 selected attributes plus Bloody sputum	0.83	0.941	0.957	0.051	0.949	0.97	94.85%
J48 with 8 selected attributes plus Headache	**0.2**	0.933	0.961	0.052	0.947	0.968	94.61%
J48 with 8 selected attributes plus Bloody sputum and Headache	0.52	0.941	0.961	0.048	0.951	0.97	95.08%

5.7.3. Effect of Dataset Size

The performance of all models developed using different algorithms degraded as the data size increases. The effect of dataset size in this study shows that increasing the dataset size decreases the true prediction rate, precision, ROC area, correctly classified instances and accuracy of the model, whereas incorrectly classified instances, false prediction rate and time taken to build model increased due to increasing dataset size. Figure 5.3 presents that the accuracy of predictive model.

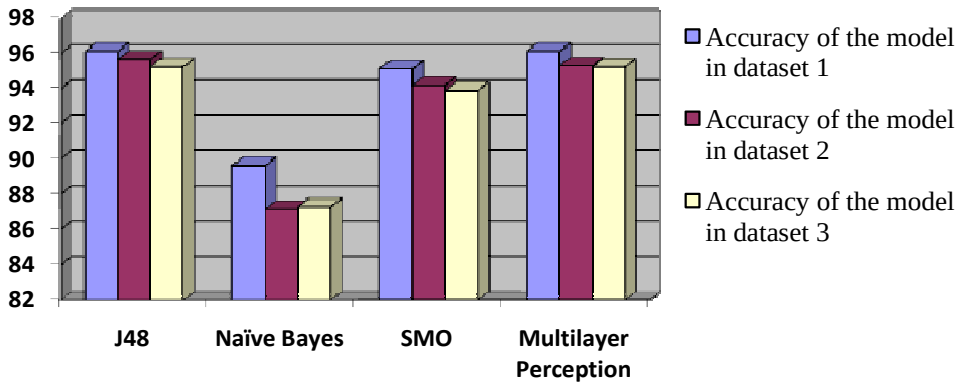


Figure 5.3 Effect of dataset size on classification accuracy.

The results of this experiment indicate that change in dataset size has impact on the performance of mining algorithms. The researcher decided to conducts additional experiment to identify the reason why the performances of mining algorithms decreased while dataset size increased. This experiment was done using J48 classifier with pruned and unpruned parameter within dataset 1, dataset 2 and dataset 3. Table 5.14 shows that the effect of pruning on the performance of predictive models as the data size increases.

Table 5.14 Effects of pruning.

	J48 with all attributes and pruned parameter			J48 with all attributes and unpruned parameter		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Dataset 1	**96.1%**	0.964	0.956	**96.1%**	0.962	0.96
Dataset 2	**95.29%**	0.944	0.963	**95.93%**	0.955	0.964
Dataset 3	**95.24%**	0.942	0.964	**95.29%**	0.945	0.962

As shown in the Table 5.14, the performances of the models declined as the dataset size increased in both pruned and unpruned models. However for dataset 2 and 3, small improvement has been observed in unpruned models compared to the pruned ones.

Based on the experiments conducted in this study, the researcher concluded that increasing dataset size decreases the performance of mining algorithms because increasing dataset size increased complexity of model, increased number of rules generated, and decrease the model accuracy, sensitivity and specificity. This is clearly illustrated in Table 5.15.

Table 5.15 Effects of dataset size J48 classifier models

J48 with all attributes	Instance Number	Number of Rules	Size of Tree (complexity)	Sensitivity	Specificity	Accuracy
Applied to dataset 1	5008	65	115	0.964	0.956	96.1%
Applied to dataset 2	7508	82	145	0.944	0.963	95.29%
Applied to dataset 3	10,031	98	177	0.942	0.964	95.24%

5.7.4. Model Comparison

One of the objectives of this study stated at the initial stage of the study was identifying which data mining algorithm performs best in predicting the TB status of patients. Therefore, the experiments were carried out in this research with decision tree, naïve bayes, support vector machine and artificial neural network algorithms. The model for each experiment was built using dataset 3 with all attributes and best selected attributes. Model comparison was performed using performance evaluation matrix like true prediction rate, false prediction rate, precision, ROC area, time taken to build the model and accuracy of model. Table 5.16 presents the result of each model.

Table 5.16 Performance evaluations of models.

Model	Time (sec)	TP Rate	TN Rate	FP Rate	Precision	ROC area	Accuracy
J48 with all attributes	0.6	0.942	**0.964**	**0.047**	**0.953**	0.972	**95.24%**
J48 with 8 selected attributes	0.5	0.933	0.957	0.054	0.945	0.968	94.44%
Naïve Bayes with all attributes	**0.01**	0.834	0.916	0.123	0.877	0.956	87.3%
Naïve Bayes with 8 selected attributes	**0.01**	0.849	0.941	0.103	0.898	0.96	89.29%
SMO with all attributes	28.4	0.945	0.932	0.062	0.939	0.939	93.89%
SMO with 8 selected attributes	9.79	0.903	0.931	0.083	0.917	0.917	91.59%
MLP with all attribute	61.37	**0.948**	0.958	**0.047**	0.952	**0.978**	95.23%
MLP with 8 selected attributes	17.22	0.936	0.955	0.054	0.945	0.975	94.46%

As shown in Table 5.16, the comparison between the models focus on which algorithms produce better outcomes. The highest accuracy of models in this study, were generated by J48 decision tree and MLP of artificial neural network. The models developed using J48 decision tree

classifiers have accuracies of 95.24% and 94.44% in experiment 1 and 2 respectively. MLP classifiers have accuracies of 95.23% and 94.46% in experiment 7 and 8 respectively.

The classifiers were also compared on the basis of time taken to build the model. Fastest classifier in the study was naïve bayes followed by J48 classifier, whereas slower data mining algorithms were SMO and MLP. The experiments conducted with selected attributes took small execution time than experiments with all attributes.

The classifiers were also compared based on sensitivity (true positive rate) and specificity (true negative rate). The highest sensitivity was scored by MLP of ANN classifier and followed by SMO of SVM and J48 decision tree classifiers, whereas lowest sensitivity was generated by naïve bayes classifier. The specificity of the developed model was high in J48 classifier followed by MLP of ANN classifier. Sensitivity and specificity of the two best performing algorithms were presented in the following two confusion matrixes (Table 5.17 and 5.18).

Table 5.17 Confusion matrix of J48 classifier with all attributes.

		Prediction Class		Total
		Negative	Positive	
Actual Class	Negative	4600 (TN)	173 (FP)	4773
	Positive	304 (FN)	4954 (TP)	5258
Total		4904	5127	10031

As it is shown in Table 5.16, in J48 classifier the correctly classified instances were 4600 + 4954 = 9554 (95.24%) and incorrectly classified instances were 304 + 173 = 477 (4.75%). The sensitivity of the model was 4954 /5258 = 0.942 and specificity of the model was 4600 / 4773 = 0.964.

Table 5.18 Confusion matrix of MLP classifier with all attributes.

		Prediction Class		Total
		Negative	Positive	
Actual Class	Negative	4571 (TN)	202 (FP)	4773
	Positive	276 (FN)	4982 (TP)	5258
Total		4847	5184	10031

As it is shown in Table 5.17, in MLP classifier the correctly classified instances were 4571 + 4982 = 9553 (95.23%) and incorrectly classified instances were 276 + 202 = 478 (4.8%). The sensitivity of the model was 4982 /5258 = 0.948 and specificity of the model was 4571 / 4773 = 0.958.

In conclusion, the best performing algorithm among the four algorithms used in this study was chosen by considering accuracy, sensitivity, specificity, false prediction rate and time taken to build model. Based on performance evaluator, the order of best algorithms begins with J48 classifier followed by MLP classifier and SMO classifier. The least performed algorithm was naïve bayes. The best algorithm selected for the given dataset was J48 decision tree.

5.8. Generated Rules from Decision Trees

In this study, J48 classifier has achieved relatively the highest in most of performance evaluation criteria compared to MLP, SMO and naïve bayes algorithms. Therefore, the model generated by J48 classifier with all (15) attributes was selected as the best model that can predict the TB status of patients'. As can be seen in appendix 4, J48 decision tree generated 98 rules for predicting TB status of patients from which the researcher considered and discussed the rules that predict at least 250 cases correctly. Based on this assumption ten rules are selected from decision tree and all of them were acceptable by the domain experts. These rules are listed and discussed below.

Rule 1. IF Shortness of breath = No AND Chest pain = No AND Bloody sputum = No AND Weight loss=No : TB status = Negative (2640.0 / 67.0)

The above rule gives correct result for 2640 (97.5%) of the 2707 cases it covers. Patients who do not have shortness of breath, chest pain, bloody sputum and weight loss are likely to be Negative in their TB status.

Rule 2. IF Shortness of breath = No AND Chest pain = No AND Bloody sputum = No AND Weight loss = No AND Night sweats = No: TB status = Negative (399.0 / 32.0)

The second rule gives correct result for 399 (92.6%) out of 431 cases. Patients not having shortness of breath, chest pain, bloody sputum, weight loss and night sweats are Negative in their TB status.

Rule 3. IF Shortness of breath = No AND Chest pain = Yes AND Cough = Yes AND Night sweats = Yes AND Loss of appetite = Yes AND Weight loss = Yes: TB status = Positive (1218.0 / 6.0)

The third rule states that the patients having chest pain, cough, night sweats, loss of appetite and weight loss are Positive in their TB status. This rule produced correct results for 1218 (99.5%) out of 1224 cases.

Rule 4. IF Shortness of breath = No AND Chest pain = Yes AND Cough = Yes AND Night sweats = No AND HIV test result = Reactive AND Weight loss = No: TB status = Negative (319.0 / 41.0)

This rule states that patients having chest pain and cough but do not have shortness of breath, night sweats and weight loss with reactive HIV test result are Negative in their TB status. This rule produced correct results for 319 (88.6%) out of 360 cases.

Rule 5. IF Shortness of breath = No AND Chest pain = Yes AND Cough = No AND Night sweats = Yes AND Weight = 51 – 70: TB status = Negative (2150.0 / 22.0)

The above rule gives correct results for 2150 (98.7%) of the 2172 cases. Patients whose weight was between 51 – 70 with chest pain and night sweats and not having shortness of breath and cough are Negative in their TB status.

Rule 6. IF Shortness of breath = No AND Chest pain = Yes AND Cough = No AND Night sweats = No: TB status = Negative (501.0 / 10.0)

This rule states that patients not having shortness of breath, cough and night sweats but having chest pain are Negative in their TB status. This rule produced correct results for 501 (98.2%) out of 511 cases.

Rule 7. IF Shortness of breath = Yes AND Cough = Yes AND Chest pain = No AND HIV test result = Reactive AND Fever = Yes AND Weight loss = No AND Headache = No: TB status = Positive (356.0 / 9.0)

The seventh rule states that patients with reactive HIV test result and having shortness of breath, cough and fever, but not having chest pain and weight loss, the result is TB Positive for 356 (97.5) cases out of 365.

Rule 8. IF Shortness of breath = Yes AND Cough = Yes AND Chest pain = No AND HIV test result = Reactive AND Fever = Yes AND Weight loss = Yes: TB status = Positive (281.0 / 1.0)

The above rule states that patients with reactive HIV test result and having shortness of breath, cough, fever, and weight loss, but does not have chest pain are Positive in their TB status. This rule gives correct result for 281 (99.6%) cases out of 282.

Rule 9. IF Shortness of breath = Yes AND Cough = Yes AND Chest pain = Yes AND Loss of appetite = No AND Night sweats = Yes: TB status = Positive (438.0 / 12.0)

The above rule states that patients having shortness of breath, cough, chest pain and night sweats, but does not have loss of appetite are Positive in their TB status. This rule produced correct results for 438 (97.3%) out of 450 cases.

Rule 10. IF Shortness of breath = Yes AND Cough = Yes AND Chest pain = Yes AND Loss of appetite = Yes: TB status = Positive (1868.0 / 10.0)

The tenth rule states that patients having shortness of breath, cough, chest pain and loss of appetite are Positive in their TB status. This rule gives correct results for 1868 (99.4%) cases out of 1878 cases.

5.9. Graphical User Interface Development

The development of graphical user interface in this study was done using Microsoft visual studio 2008. This prototype graphical user interface is developed based on the model generated by J48 decision tree classifier with pruned parameter and all attributes. The rules used by the researcher to design the graphical user interface for predicting the TB status of patients are the 10 rules listed and briefly discussed in section 5.8. Figure 5.4 shows the prototype user interface developed for predicting the TB status.

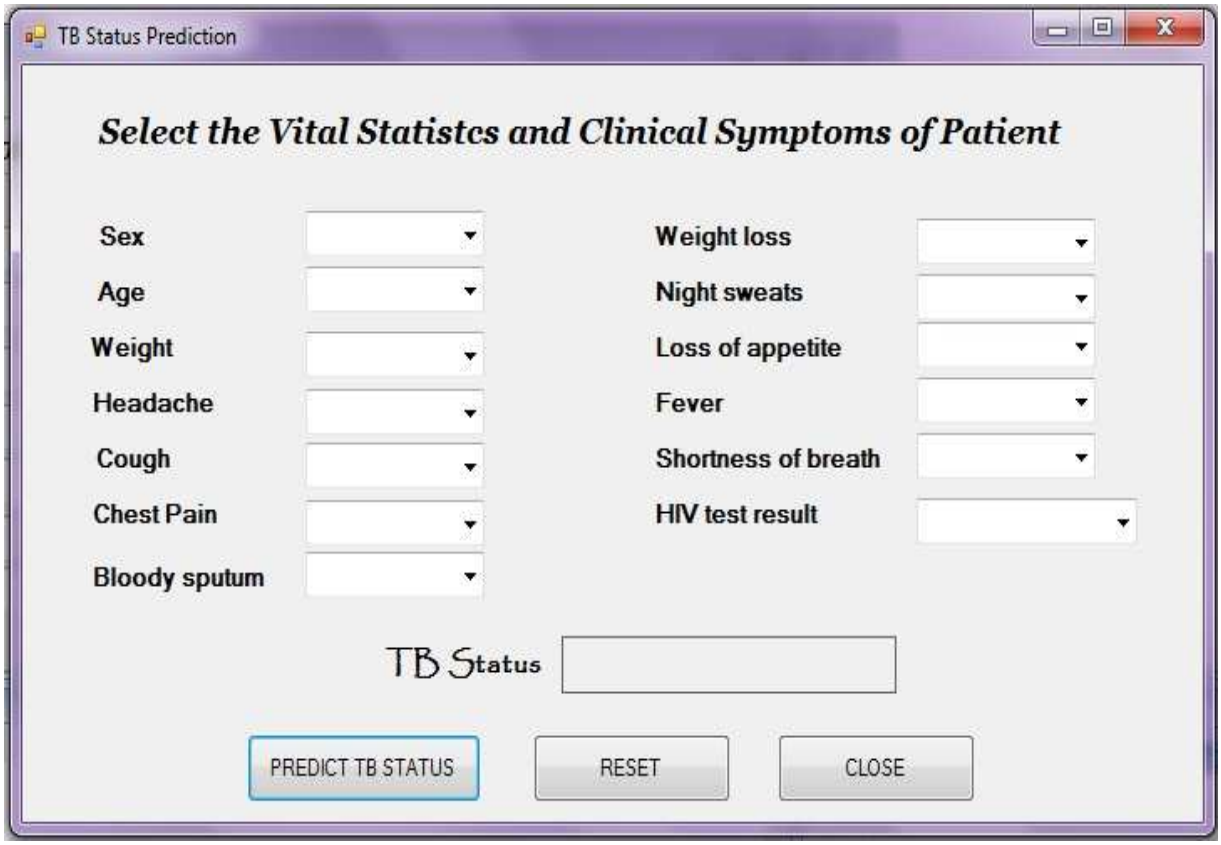


Figure 5.4 Prototype of the user interface for TB status prediction

As shown in Figure 5.5, this prototype prediction model can be used for predicting TB status of patients based on the rules generated by J48 classifier. The sample result on Figure 6.5 shows that the prototype predicts the TB status of the patient as positive based on rule 7 shown in section 5.8. The visual basic codes to predict TB status of patients written based on the ten selected rules generated by J48 decision tree are presented in appendix 5 and 6.

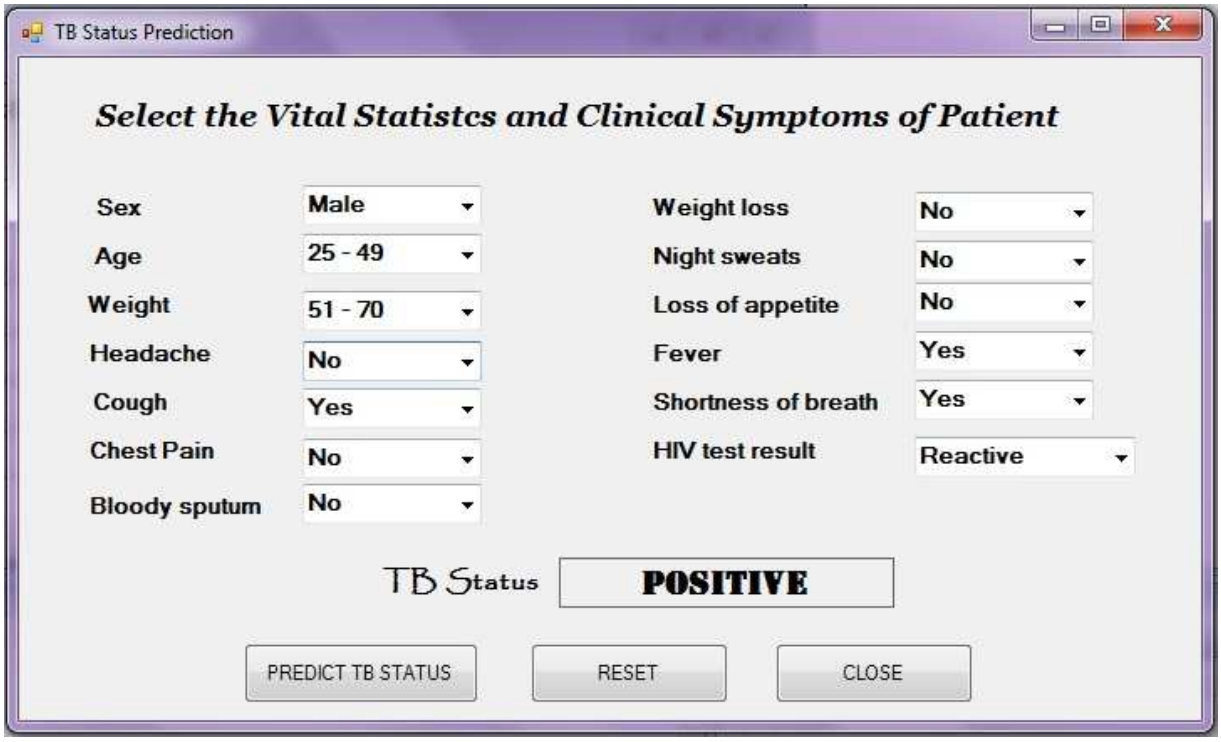


Figure 5.5 TB status prediction prototype user interface with sample result

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

Tuberculosis is a disease affecting mostly young adults in their most productive years. In Ethiopia, TB is a disease of major public health problem. Ethiopia is now ranked seventh of the 22 highest TB burden countries, which collectively contribute 80% of the global TB burden. Early identification and isolation of TB cases is critical to prevent further transmission, morbidity and mortality caused by TB. Data mining (DM) has a potential to indentify hidden knowledge from huge datasets. Many researchers apply data mining to explore hidden pattern from electronic medical records. It is possible to use DM algorithms for analysis and predicting the prevalence of positive TB cases.

The goal of this research was to apply the data mining techniques for predicting the TB status of patients from patients' dataset. This study attempted to identify the determinant attributes affecting TB status of a patient, extract useful predictive model from patients' dataset and developed a prototype graphical user interface for effective utilization of the predictive model.

In this research, the methodology was hybrid data mining process model, that involves six steps and the researcher properly went through all steps. This study considers a total of 10,031 records and 15 attributes to predict the TB status. Handling of missing values and data transformation was done to prepare the dataset for experiments. The mining algorithms were used in order to build the models that can predict the TB status of the patients'. The algorithms used are decision tree, naïve bayes, support vector machine and artificial neural network. To evaluate the model accuracy 10-fold cross validation was used and confusion matrix was used to evaluate model performance.

Experiments are conducted to evaluate the performance of predictive models as the size of the dataset increases. The results of this experiment shows that increasing dataset size decrease the performance of mining algorithms because increasing dataset size increased complexity of model, increase number of rules generated, and decrease the model accuracy, sensitivity and specificity.

The prediction models were developed with all (15) attributes and 8 best attributes selected by information gain attribute evaluator. The performance of J48, SMO and MLP was better with all attributes than selected best attributes, whereas in naïve bayes classifier, the performance increased in selected attributes.

The most important attributes that determine TB states of the patients are shortness of breath, chest pain, cough, weight loss, loss of appetite, night sweats and HIV test result.

The evaluation of best performed algorithms compared based on accuracy, sensitivity, specificity, false prediction rate and time taken to build model. Based on performance evaluator, the best performing algorithms are J48 classifier followed by MLP classifier and SMO classifier. The least performed algorithm was naïve bayes classifier. The best selected model in this study was generated by J48 decision tree with all attributes. The accuracy of this model is 95.24%. Hence, the best algorithm selected for the give dataset was J48 decision tree.

J48 classifier generated model was selected as the best model that can predict the TB status of patients. J48 decision tree generated 98 rules for predicting TB status of the patients from which the researcher considered only 10 rules that predict at least 250 TB cases correctly. Graphical user interface was designed using these ten rules. The development of graphical user interface in this study was done using Microsoft visual studio 2008.

The results achieved from this research indicate that data mining is useful in bringing relevant information from large and complex dataset so that anybody can use this information for decision making.

6.2 Recommendation

In this research, an attempt has been made to find out the potential applicability of data mining technology to predict the prevalence of Positive TB cases. Even if this research has been done for the academic exercise, its results are promising to be applied in addressing practical problems in prevention and control of tuberculosis. The researcher forwarded the following recommendations based on the result of this study:

- This research attempted to develop prediction model and prototype graphical interface for TB status prediction. There is, however, a need for the development of knowledge based system for TB status prediction with domain experts. This should be a future research direction.

- Decision tree, naïve bayes, support vector machine and artificial neural network were the selected mining algorithms for model development in this study. However, experiments using more machine learning algorithms along with much large data size need to be conducted to see the effect and optimize the prediction of TB status.

- The study areas for this research are Menelik II and St. Peters TB specialized hospitals. These two hospitals were using paper based patient medical records when this research was conducted. This made the research difficult and time consuming to encode important variables from paper based medical records to computer. Therefore, the implementation of electronic medical record would be helpful for application of data mining in the domain area.

- In this study, experiments are conducted using J48, naïve bayes, support vector machine and artificial neural network with default parameter of WEKA software. Therefore, further researches can be conducted by changing parameters of these four algorithms to boost the accuracy of the models.

- Initially, the researcher planned to conduct this research on the basis of the previous research by Asia Nusredin. Accordingly, the researcher tried to get Menelik II patients' dataset that was already used by the previous researcher. But it was not possible to get that dataset from the school of information science and the previous researcher. As a

result, the researcher was forced to collect and encode patients' data from paper based medical records. To alleviate the problems new researchers face in this area, the school and other concerned bodies should collect both the report work and dataset used for model development. Researchers should also submit their research works and dataset for concerned stakeholders of the research.

REFERENCE

1. Vincent H. Developing a consumer health informatics decision support system using concept analysis; 2007.

2. National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (CDC HIV), online accessed from www.cdc.gov/hiv. Accessed date on November 29, 2012 at 6:00 PM.

3. WHO (2010/2011), Tuberculosis Global Facts, online accessed from www.who.int/tb/data. Accessed date on November 27, 2012 at 4:00 PM.

4. US Global Health Police, Tuberculosis High-Burden Countries (HBCs) 2012, online accessed from http://www.globalhealthfacts.org/data/topic/map.aspx?ind=111 . Accessed date on December 1, 2012 at 4:00 PM.

5. Tamer U. Predicting the Existence of Mycobacterium Tuberculosis on Patients by Data Mining Approach, School of Natural and Applied Sciences, Computer Engineering, Bahcesehir University, Istanbul, Turkey; 2009.

6. Federal Ministry of Health, Disease prevention and control program: ADIS in Ethiopia, 3rd ed., Addis Ababa, Ethiopia; 2000.

7. Federal Ministry of Health, Tuberculosis, Leprosy and TB/HIV prevention and control program manual, 4th ed., Addis Ababa, Ethiopia; 2008.

8. Office of Technology Assessment, United States Congress, The Continuing Challenge of Tuberculosis; 1993.

9. Asha.T,S. Natarajan, K.N.B. Murthy. Estimating the Statistical Significance of Classifiers used in the Prediction of Tuberculosis. Journal of Computer Engineering [Internet]. Vol. 5, PP 05-13; 2012: 05-13

10. Ali A.E, Chiu-Bin Hsiao, Susan Goodnough, Joseph Serghani, Brydon J. B. Grant. Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network [Internet]. 1999; 116: 968–973

11. Asia Nusredin. Mining patients' data for effective tuberculosis diagnosis: The case of Menelik II hospital. Adiss Ababa university; 2012.

12. Young D.B. Future Research Needs in Tuberculosis, London, UK; 1994.

13. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Inc; 2003.

14. Sumathi S, Sivanandam S.N. Introduction to Data Mining and its Applications, Springer-Verlag Berlin Heidelberg; 2006.

15. Jiawei Han, Micheline Kamber. Data Mining: concept and techniques, Simon Fraser university, Morgan Kaufmann; 2001.

16. Fayyad U, Piatesky-Shapiro G, Smyth P. Knowledge Discovery and Data mining: Towards a Unifying Framework. CA; 1996.

17. Witten I.H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd ed; 2005.

18. Berry M, Linoff G. Data Mining Techniques: For marketing, sales and customer support. New York: John Wiley and Sons, Inc; 1997.

19. David H, Heikki M, Padhraic S. Principles of Data mining, Massachusetts, London, England: MIT press; 2001.

20. Daniel T. Larose. Discovering Knowledge in Data: An Introduction to Data Mining, A John Wiley & Sons, INC.; 2005.

21. Lior Rokach. Decomposition Methodology in Data Mining. Department of Information Systems Engineering, Faculty of Engineering Sciences: Ben-Gurion University of the Negev; 2006.

22. Daniel T. Larose. Discovering Knowledge in Data: An Introduction to Data Mining, A John Wiley & Sons, INC.; 2005.

23. Cios K., Witold P. Roman S. and Kurgan A. Data Mining: A Knowledge Discovery Approach, New York, USA: Springer; 2007.

24. Mitchell M. Machine Learning. New York: The McGraw-Hill Companies, Inc; 1997.

25. Haykin S. Neural Networks: A comprehensive Foundation. USA. Prentice-Hall, inc; 1994.

26. Popescu M., Perscu-Popesu L., Valentina E., Balas V. E. and Mastorakis N. Multilayer Perceptron and Neural Networks; 2009.

27. Naren R, David A. H, Benjamin J. K. Mining Electronic Health Records, 2010; 95-99

28. Fox GJ, Dobler C.C, Marks GB. Active case finding in contacts of people with tuberculosis (Review), Cochrane Collaboration and published in the Cochrane Library; 2011.

29. WHO. TB/ HIV: Clinical manual, Stop TB Department, 2nd ed., Geneva; 2004.

30. WHO, Global Tuberculosis Control Report 2010. Geneva, Switzerland: WHO; 2010.

31. Mitchell M., Machine Learning. New York: The McGraw-Hill Companies, Inc; 1997.

32. WHO, THE GLOBAL PLAN TO STOP TB 2006-2015: Stop TB Strategy. Geneva, Switzerland: WHO; 2005.

33. FMOH, Ethiopia Tuberculosis, Leprosy and TB/HIV Prevention and Control Programme, Manual, 4th Ed; 2008.

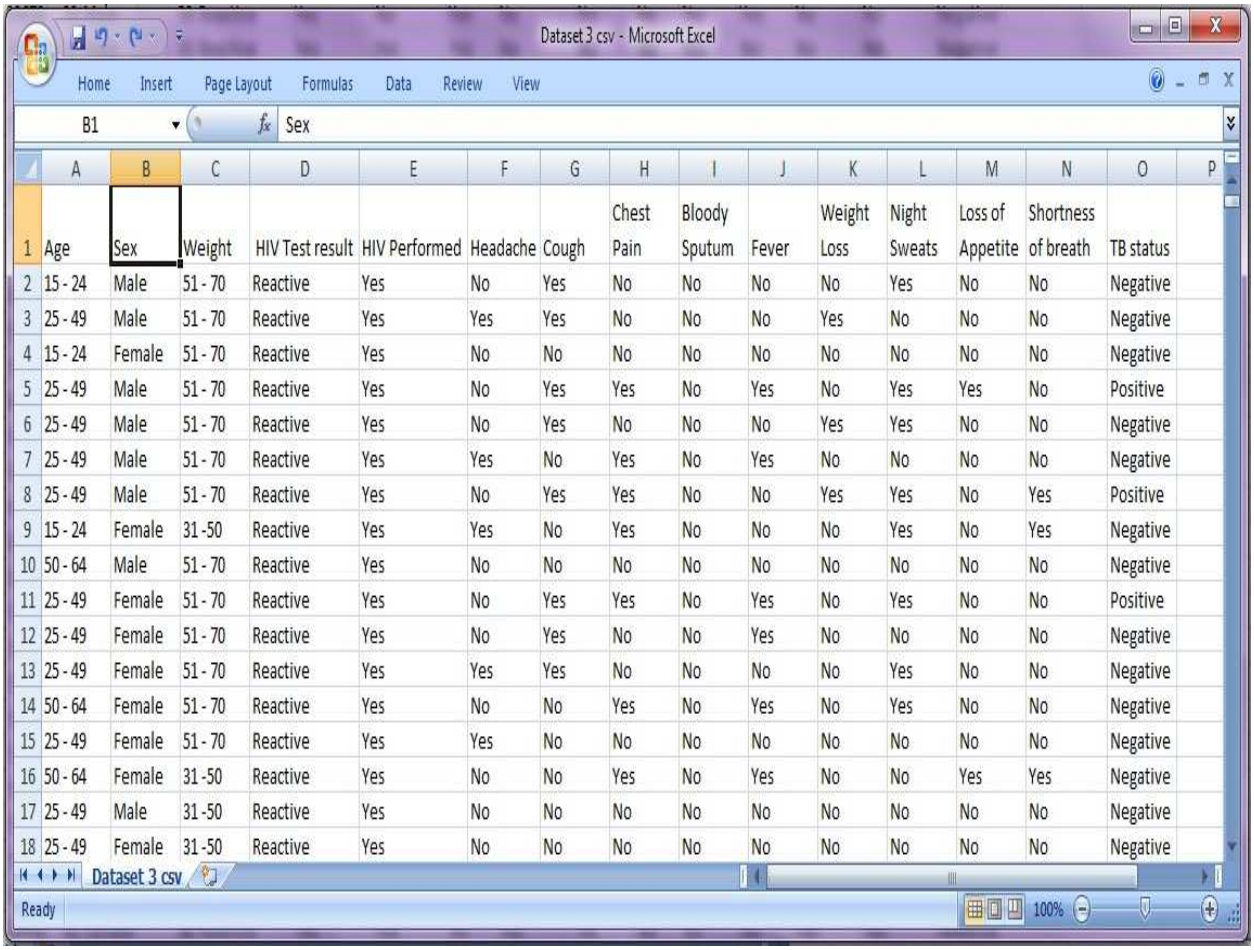
34. WHO, Global Tuberculosis Control Report 2011. Geneva, Switzerland: WHO; 2011.

35. FMOH, Guidelines for clinical and programmatic management of TB, Leprosy and TB/HIV in Ethiopia, Addis Ababa, Ethiopia, 5th Ed; 2012.

36. Addis Ababa City Administration, online accessed from http://www.addisababacity.gov.et. Accessed date on January 3, 2013 at 7:16 PM.

APPENDIXES

Appendix 1: Dataset Sample with CSV (comma delimited) File Format



Dataset 3 csv - Microsoft Excel

Home Insert Page Layout Formulas Data Review View

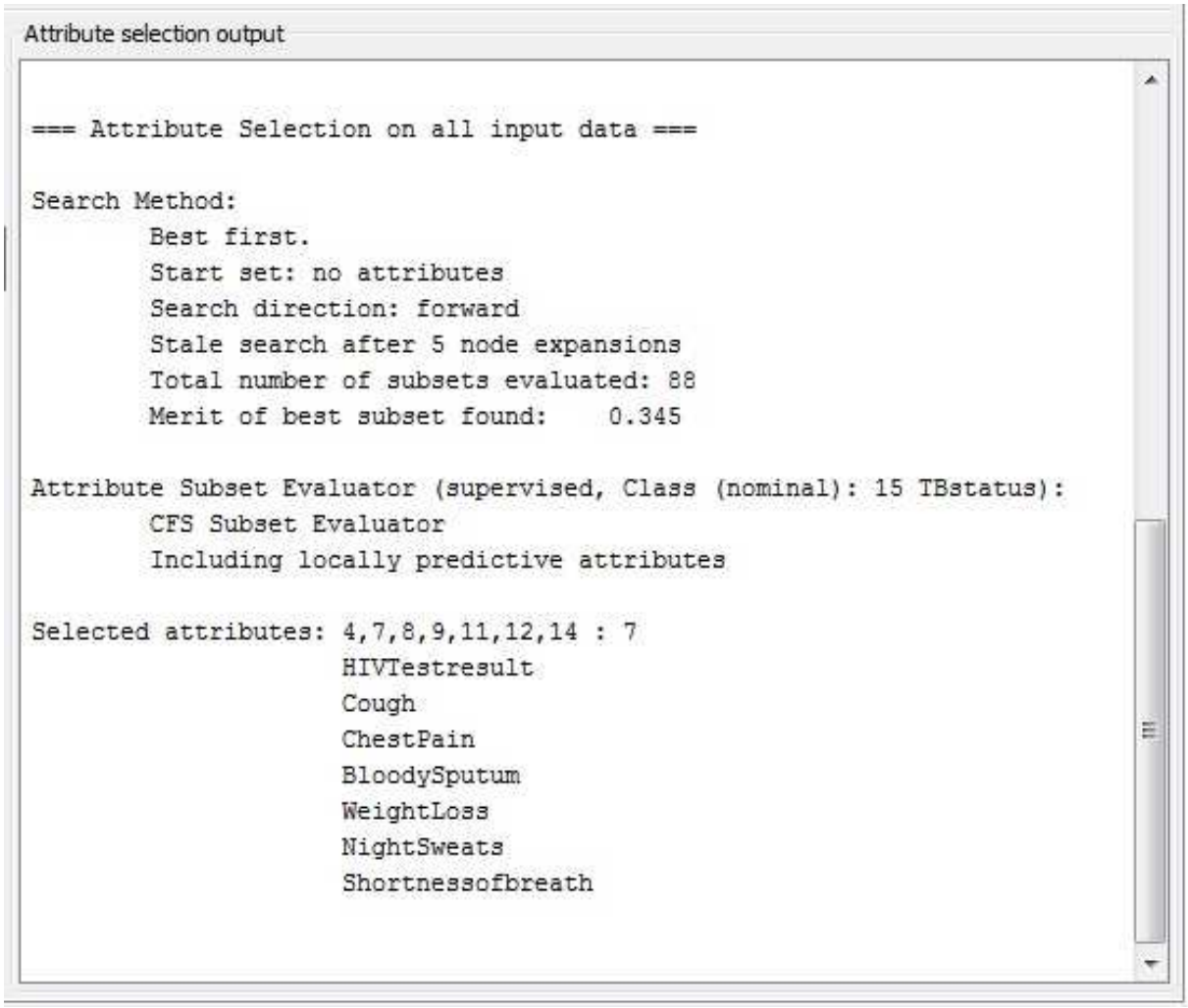
B1 Sex

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Age	Sex	Weight	HIV Test result	HIV Performed	Headache	Cough	Chest Pain	Bloody Sputum	Fever	Weight Loss	Night Sweats	Loss of Appetite	Shortness of breath	TB status	
2	15 - 24	Male	51 - 70	Reactive	Yes	No	Yes	No	No	No	No	Yes	No	No	Negative	
3	25 - 49	Male	51 - 70	Reactive	Yes	Yes	Yes	No	No	No	Yes	No	No	No	Negative	
4	15 - 24	Female	51 - 70	Reactive	Yes	No	No	No	No	No	No	No	No	No	Negative	
5	25 - 49	Male	51 - 70	Reactive	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Positive	
6	25 - 49	Male	51 - 70	Reactive	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Negative	
7	25 - 49	Male	51 - 70	Reactive	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Negative	
8	25 - 49	Male	51 - 70	Reactive	Yes	No	Yes	Yes	No	No	Yes	Yes	No	Yes	Positive	
9	15 - 24	Female	31 -50	Reactive	Yes	Yes	No	Yes	No	No	No	Yes	No	Yes	Negative	
10	50 - 64	Male	51 - 70	Reactive	Yes	No	No	No	No	No	No	No	No	No	Negative	
11	25 - 49	Female	51 - 70	Reactive	Yes	No	Yes	Yes	No	Yes	No	Yes	No	No	Positive	
12	25 - 49	Female	51 - 70	Reactive	Yes	No	Yes	No	No	Yes	No	No	No	No	Negative	
13	25 - 49	Female	51 - 70	Reactive	Yes	Yes	Yes	No	No	No	No	Yes	No	No	Negative	
14	50 - 64	Female	51 - 70	Reactive	Yes	No	No	Yes	No	Yes	No	Yes	No	No	Negative	
15	25 - 49	Female	51 - 70	Reactive	Yes	Yes	No	No	No	No	No	No	No	No	Negative	
16	50 - 64	Female	31 -50	Reactive	Yes	No	No	Yes	No	Yes	No	No	Yes	Yes	Negative	
17	25 - 49	Male	31 -50	Reactive	Yes	No	No	No	No	No	No	No	No	No	Negative	
18	25 - 49	Female	31 -50	Reactive	Yes	No	No	No	No	No	No	No	No	No	Negative	

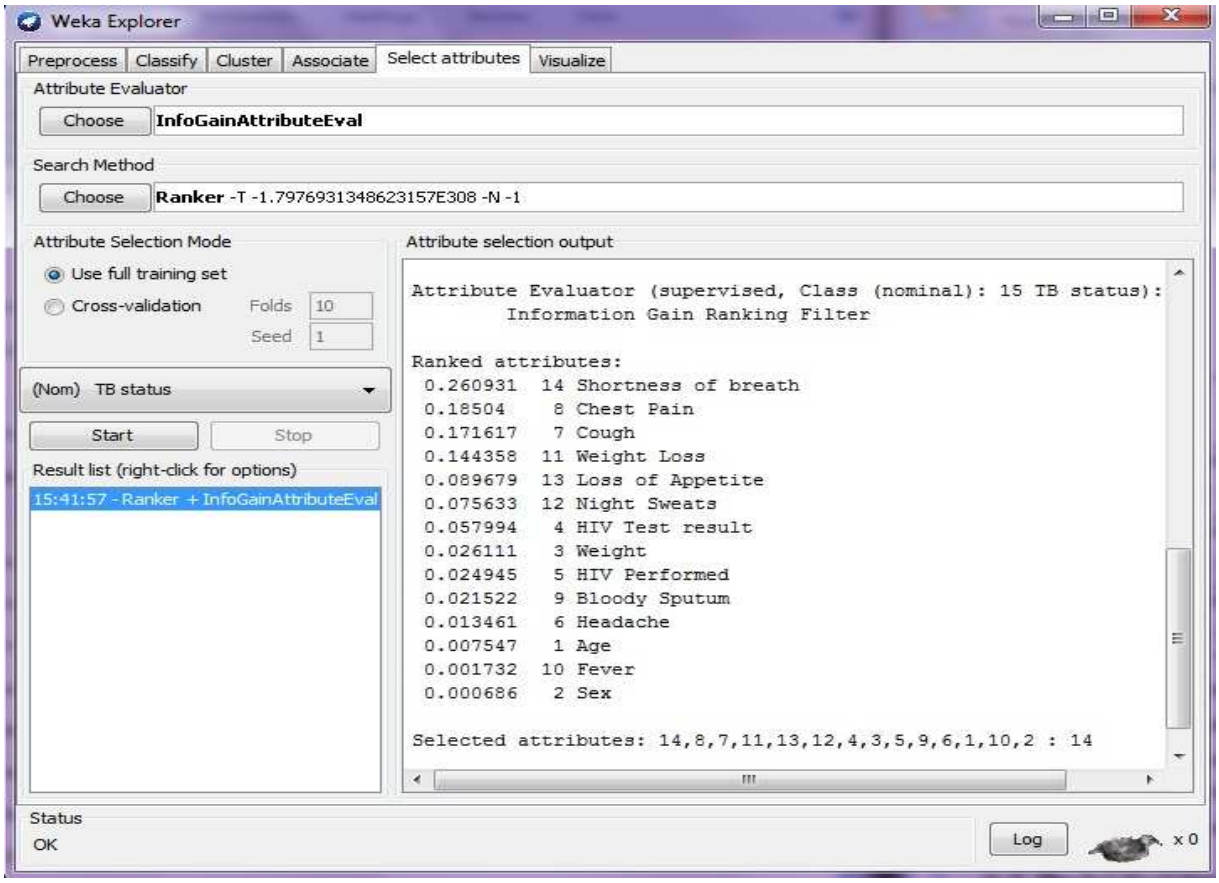
Dataset 3 csv

Ready 100%

Appendix 2: Result of CFS Attributes Subset Evaluator



Appendix 3: Result of Information gain Attribute Evaluator



Appendix 4: Rules generated by J48 decision tree

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Dataset 3 csv

Instances: 10031

Attributes: 15

Age
Sex
Weight
HIV Test result
HIV Performed
Headache
Cough
Chest Pain
Bloody Sputum
Fever
Weight Loss
Night Sweats
Loss of Appetite
Shortness of breath
TB status

Test mode: 10-fold cross-validation

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

J48 pruned tree

Shortness of breath = No

| Chest Pain = No

| | Bloody Sputum = No

| | | Weight Loss = No: Negative (2640.0/67.0)

| | | Weight Loss = Yes

| | | | Night Sweats = Yes

| | | | | Loss of Appetite = No: Negative (67.0/12.0)

| | | | | Loss of Appetite = Yes

| | | | | | Cough = Yes: Positive (58.0/21.0)

| | | | | | Cough = No

| | | | | | | Sex = Male: Negative (4.0)

| | | | | | | Sex = Female: Positive (3.0/1.0)

| | | | Night Sweats = No: Negative (399.0/32.0)

| | Bloody Sputum = Yes

| | | Cough = Yes

| | | | Weight Loss = No

| | | | | Fever = No

| | | | | | Headache = No: Negative (31.0/4.0)

| | | | | | Headache = Yes

| | | | | | | Weight = 51 - 70

| | | | | | | | HIV Performed = Yes: Positive (4.0)

| | | | | | | | HIV Performed = No: Negative (3.0)

| | | | | | | Weight = 31 -50: Negative (5.0)

| | | | | | | Weight = Below 30: Negative (0.0)

| | | | | | | Weight = Above 70: Negative (0.0)

| | | | | Fever = Yes: Positive (13.0/2.0)

| | | | Weight Loss = Yes

| | | | | HIV Performed = Yes: Positive (34.0/1.0)

| | | | | HIV Performed = No

| | | | | | Night Sweats = Yes: Positive (6.0)

| | | | | | Night Sweats = No: Negative (2.0)

| | | Cough = No: Negative (57.0/2.0)

| Chest Pain = Yes

| | Cough = Yes

| | | Night Sweats = Yes

| | | | Loss of Appetite = No

| | | | | HIV Test result = Reactive

| | | | | | Weight Loss = No

| | | | | | | Bloody Sputum = No

| | | | | | | | Headache = No: Negative (54.0/8.0)

| | | | | | | | Headache = Yes

| | | | | | | | | Weight = 51 - 70: Negative (8.0/2.0)

| | | | | | | | | Weight = 31 -50: Positive (3.0)

| | | | | | | | | Weight = Below 30: Positive (1.0)

| | | | | | | | | Weight = Above 70: Positive (1.0)

| | | | | | | Bloody Sputum = Yes: Positive (5.0)

| | | | | | Weight Loss = Yes: Positive (23.0/3.0)

| | | | | HIV Test result = Non Reactive

| | | | | | Fever = No: Negative (5.0/2.0)

| | | | | | Fever = Yes: Positive (149.0/3.0)

| | | | | HIV Test result = Not Tested

| | | | | | Weight Loss = No

| | | | | | | Bloody Sputum = No: Negative (63.0/8.0)

| | | | | | | Bloody Sputum = Yes: Positive (6.0)

| | | | | | Weight Loss = Yes: Positive (14.0/1.0)

| | | | Loss of Appetite = Yes

| | | | | Weight Loss = No

| | | | | | Fever = No

| | | | | | | Bloody Sputum = No

| | | | | | | | HIV Test result = Reactive

| | | | | | | | | Sex = Male: Positive (12.0/3.0)

| | | | | | | | | Sex = Female: Negative (13.0/3.0)

| | | | | | | | HIV Test result = Non Reactive: Positive (5.0)

| | | | | | | | HIV Test result = Not Tested

| | | | | | | | | Weight = 51 - 70: Negative (12.0/3.0)

| | | | | | | | | Weight = 31 -50: Positive (14.0/5.0)

| | | | | | | | | Weight = Below 30: Positive (2.0)

| | | | | | | | | Weight = Above 70: Negative (1.0)

| | | | | | | | Bloody Sputum = Yes: Positive (7.0)

| | | | | | | Fever = Yes: Positive (195.0/11.0)

| | | | | | Weight Loss = Yes: Positive (1218.0/6.0)

| | | Night Sweats = No

| | | | HIV Test result = Reactive

| | | | | Weight Loss = No: Negative (319.0/41.0)

| | | | | Weight Loss = Yes

| | | | | | Loss of Appetite = No: Negative (12.0/2.0)

| | | | | | Loss of Appetite = Yes: Positive (25.0/5.0)

| | | | HIV Test result = Non Reactive

| | | | | Headache = No

| | | | | | Weight Loss = No: Negative (21.0/2.0)

| | | | | | Weight Loss = Yes

| | | | | | | Fever = No: Negative (4.0/1.0)

| | | | | | | Fever = Yes: Positive (3.0)

| | | | | Headache = Yes

| | | | | | Fever = No: Positive (157.0/4.0)

| | | | | | Fever = Yes

| | | | | | | Bloody Sputum = No: Negative (5.0)

| | | | | | | Bloody Sputum = Yes: Positive (2.0)

| | | | HIV Test result = Not Tested

| | | | | Bloody Sputum = No

| | | | | | Weight Loss = No: Negative (173.0/8.0)

| | | | | | Weight Loss = Yes

| | | | | | | Headache = No

| | | | | | | | Fever = No: Negative (5.0/1.0)

| | | | | | | | Fever = Yes

| | | | | | | | | Loss of Appetite = No: Negative (2.0)

| | | | | | | | | Loss of Appetite = Yes: Positive (5.0)

| | | | | | | | Headache = Yes: Negative (4.0)

| | | | | Bloody Sputum = Yes: Positive (5.0)

| | Cough = No

| | | Night Sweats = Yes

| | | | Weight = 51 - 70: Negative (215.0/22.0)

| | | | Weight = 31 -50

| | | | | Weight Loss = No

| | | | | | Age = 15 - 24: Negative (24.0/3.0)

| | | | | | Age = 25 - 49

| | | | | | | Fever = No: Negative (9.0/1.0)

| | | | | | | Fever = Yes

| | | | | | | | HIV Test result = Reactive

| | | | | | | | | Headache = No: Negative (3.0/1.0)

| | | | | | | | | Headache = Yes: Positive (17.0/1.0)

| | | | | | | | HIV Test result = Non Reactive: Positive (0.0)

| | | | | | | | HIV Test result = Not Tested: Negative (2.0)

| | | | | | Age = 50 - 64: Positive (5.0/1.0)

| | | | | | Age = Above 65: Negative (0.0)

| | | | | | Age = 0 - 14: Negative (0.0)

| | | | | Weight Loss = Yes: Positive (4.0)

| | | | Weight = Below 30: Negative (8.0/1.0)

| | | | Weight = Above 70: Negative (7.0)

| | | Night Sweats = No: Negative (501.0/10.0)

Shortness of breath = Yes

| Cough = Yes

| | Chest Pain = No

| | | HIV Test result = Reactive

| | | | Fever = No

| | | | | Night Sweats = Yes: Positive (53.0/8.0)

| | | | | Night Sweats = No

| | | | | | Bloody Sputum = No: Negative (13.0/2.0)

| | | | | | Bloody Sputum = Yes: Positive (7.0)

| | | | Fever = Yes

| | | | | Weight Loss = No

| | | | | | Headache = No: Positive (356.0/9.0)

| | | | | | Headache = Yes

| | | | | | | Night Sweats = Yes: Positive (21.0/5.0)

| | | | | | | Night Sweats = No: Negative (13.0/3.0)

| | | | | Weight Loss = Yes: Positive (281.0/1.0)

| | | HIV Test result = Non Reactive

| | | | Fever = No: Negative (11.0/2.0)

| | | | Fever = Yes

| | | | | Night Sweats = Yes: Positive (11.0)

| | | | | Night Sweats = No: Negative (4.0/1.0)

| | | HIV Test result = Not Tested

| | | | Night Sweats = Yes

| | | | | Weight Loss = No

| | | | | | Bloody Sputum = No

| | | | | | | Loss of Appetite = No

| | | | | | | | Fever = No: Negative (11.0)

| | | | | | | | Fever = Yes

| | | | | | | | | Sex = Male: Positive (3.0)

| | | | | | | | | Sex = Female: Negative (4.0/1.0)

| | | | | | | Loss of Appetite = Yes: Positive (14.0/2.0)

| | | | | | Bloody Sputum = Yes: Positive (9.0)

| | | | | Weight Loss = Yes: Positive (23.0)

| | | | Night Sweats = No: Negative (27.0/4.0)

| | Chest Pain = Yes

| | | Loss of Appetite = No

| | | | Night Sweats = Yes: Positive (438.0/12.0)

| | | | Night Sweats = No

| | | | | Weight Loss = No: Negative (27.0/9.0)

| | | | | Weight Loss = Yes: Positive (11.0/2.0)

| | | Loss of Appetite = Yes: Positive (1868.0/10.0)

| Cough = No

| | Bloody Sputum = No

| | | Weight Loss = No: Negative (116.0/8.0)

| | | Weight Loss = Yes

| | | | Night Sweats = Yes

| | | | | Loss of Appetite = No: Negative (2.0)

| | | | | Loss of Appetite = Yes: Positive (5.0)

| | | | Night Sweats = No

| | | | | Fever = No: Negative (10.0)

| | | | | Fever = Yes

| | | | | | Weight = 51 - 70: Negative (8.0/2.0)

| | | | | | Weight = 31 -50: Positive (2.0)

| | | | | | Weight = Below 30: Negative (0.0)

| | | | | | Weight = Above 70: Negative (0.0)

| | Bloody Sputum = Yes: Positive (9.0)

Number of Leaves : 98

Size of the tree : 177

Time taken to build model: 0.6 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	9554	95.2447 %
Incorrectly Classified Instances	477	4.7553 %
Kappa statistic	0.9048	
Mean absolute error	0.0754	
Root mean squared error	0.2047	
Relative absolute error	15.1082 %	
Root relative squared error	40.9906 %	
Total Number of Instances	10031	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.964	0.058	0.938	0.964	0.951	0.972	Negative
	0.942	0.036	0.966	0.942	0.954	0.972	Positive
Weighted Avg.	0.952	0.047	0.953	0.952	0.952	0.972	

=== Confusion Matrix ===

a	b	<-- classified as
4600	173	| a = Negative
304	4954	| b = Positive

Appendix 5: Visual Basic Code for Reset and Close Buttons

```
Private Sub Button3_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Button3.Click, Button3.Click
    End
End Sub
```

```
Private Sub Button2_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Button2.Click
    cb1.Text = ""
    cb2.Text = ""
    cb3.Text = ""
    cb4.Text = ""
    cb5.Text = ""
    cb6.Text = ""
    cb7.Text = ""
    cb8.Text = ""
    cb9.Text = ""
    cb10.Text = ""
    cb11.Text = ""
    cb12.Text = ""
    cb13.Text = ""
    tbstatus.Text = ""
End Sub
```

Appendix 6: Visual Basic Code to Predict TB Status Button

```
Private Sub Button1_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Button1.Click
    If cb12.Text = "No" And cb6.Text = "No" And cb7.Text = "No" And cb8.Text = "No" Then
        tbstatus.Text = "Negative"
    ElseIf cb12.Text = "No" And cb6.Text = "No" And cb7.Text = "No" And cb8.Text = "Yes" And cb9.Text = "No" Then
        tbstatus.Text = "Negative"
    ElseIf cb12.Text = "No" And cb6.Text = "Yes" And cb5.Text = "Yes" And cb9.Text = "Yes" And cb10.Text = "Yes" And cb8.Text = "Yes" Then
        tbstatus.Text = "Positive"
    ElseIf cb12.Text = "No" And cb6.Text = "Yes" And cb5.Text = "Yes" And cb9.Text = "No" And cb13.Text = "Reactive" And cb8.Text = "No" Then
        tbstatus.Text = "Negative"
    ElseIf cb12.Text = "No" And cb6.Text = "Yes" And cb5.Text = "No" And cb9.Text = "Yes" And cb3.Text = "51 - 70" Then
        tbstatus.Text = "Negative"
    ElseIf cb12.Text = "No" And cb6.Text = "Yes" And cb5.Text = "No" And cb9.Text = "No" Then
        tbstatus.Text = "Negative"
    ElseIf cb12.Text = "Yes" And cb5.Text = "Yes" And cb6.Text = "No" And cb13.Text = "Reactive" And cb11.Text = "Yes" And cb8.Text = "No" And cb4.Text = "No
        tbstatus.Text = "Positive"
    ElseIf cb12.Text = "Yes" And cb5.Text = "Yes" And cb6.Text = "No" And cb13.Text = "Reactive" And cb11.Text = "Yes" And cb8.Text = "Yes" Then
        tbstatus.Text = "Positive"
    ElseIf cb12.Text = "Yes" And cb5.Text = "Yes" And cb6.Text = "Yes" And cb10.Text = "No" And cb9.Text = "Yes" Then
        tbstatus.Text = "Positive"
    ElseIf cb12.Text = "Yes" And cb5.Text = "Yes" And cb6.Text = "Yes" And cb10.Text = "Yes" Then
        tbstatus.Text = "Positive"
    ElseIf cb1.Text = "" And cb1.Text = "" And cb2.Text = "" And cb3.Text = "" And cb4.Text = "" And cb5.Text = "" And cb6.Text = "" And cb7.Text = "" And cb
        tbstatus.Text = "Select Data First"
    Else
        tbstatus.Text = "Unknown Case"
    End If
End Sub
```